

# Conditional density estimation in a censored single-index regression model

OLIVIER BOUAZIZ\* and OLIVIER LOPEZ

Laboratoire de Statistique Théorique et Appliquée,  
Université Paris VI,  
175 rue du Chevaleret  
75013 Paris, France

## Abstract

Under a single-index regression assumption, we introduce a new semiparametric procedure to estimate a conditional density of a censored response. The regression model can be seen as a generalization of Cox regression model and also as a profitable tool to perform dimension reduction under censoring. This technique extends the results of Delecroix *et al.* (2003). We derive consistency and asymptotic normality of our estimator of the index parameter by proving its asymptotic equivalence with the (uncomputable) maximum likelihood estimator, using martingales results for counting processes and arguments of empirical processes theory. Furthermore, we provide a new adaptive procedure which allows us both to chose the smoothing parameter involved in our approach and to circumvent the weak performances of Kaplan-Meier estimator (1958) in the right-tail of the distribution. Through a simulation study, we study the behavior of our estimator for small samples.

*Keywords:* asymptotic normality; empirical processes; censoring; martingales for counting processes; pseudo-maximum likelihood; single-index model

## 1 Introduction

A major issue of recent papers dealing with censored regression is to propose alternatives to the popular Cox regression model. This model, also known as multiplicative hazard regression

model (see Cox (1972)), states some semiparametric model on the conditional hazard function. Estimation in this model is traditionally performed using pseudolikelihood techniques, and the theoretical properties of these procedures are covered by a large number of papers (see e.g. Fleming and Harrington (1991)). However, in some situations, the assumptions of Cox regression model are obviously not satisfied by the data set. In this paper, our aim is to perform estimation in a semiparametric regression model which allows more flexibility than the Cox regression model. This new technique can be seen as a particularly interesting alternative, since it is valid in a larger number of situations than the multiplicative hazard model.

Alternatives to Cox regression model mostly focus on the estimation of a conditional expectation, or of a quantile regression model. Koul *et al.* (1981), Stute (1999), Delecroix *et al.* consider mean-regression models where the regression function belongs to a parametric family, but with an unknown distribution of the residuals. Parametric quantile regression was studied by Gannoun *et al.* On the other hand, Lu and Burke (2005) and Lopez (2008) considered a semiparametric single-index regression model. Single-index regression models were initially introduced to circumvent the so-called “curse of dimensionality” in nonparametric regression (see. e.g. Ichimura (1993)), by assuming that the conditional expectation only depends on an unknown linear combination of the covariates. Another appealing aspect of such kind of models is that they include the Cox regression model as a particular case. The main assumption of this model is that the conditional density only depends on an unknown linear combination of the covariates, while the multiplicative hazard model states a similar assumption on the conditional hazard rate. In this paper we focus on estimation of the parameter in a regression model in which the conditional density of the response satisfies a single-index assumption. We provide asymptotic results for a new M-estimation procedure for the index parameter. This procedure can be seen as a generalization of the method of Delecroix *et al.* (2003) to the case of censored regression.

As in the uncensored case, we show that, regarding to the estimation of the parametric part of our model, there is an asymptotic equivalence between our semiparametric approach and a parametric one relying on some prior knowledge on the family of regression functions. For the nonparametric part, we use kernel estimators of conditional densities as in Delecroix *et al.* (2003). Since the performance of kernel estimators strongly relies on the choice of the smoothing parameter, we also provide a method to choose this parameter adaptively. Another technical issue in our approach concerns a truncation parameter involved in our procedure. This problem

of truncation directly comes from the censored framework, where estimators of the underlying distribution functions sometimes fail to estimate correctly the tail of the distribution. This problem is traditionally circumvented for example by assuming integrability assumptions on the response and censoring distribution, see e.g. Stute (1999). On the other hand, truncation procedure consists of removing the observations which are too large in the estimation of the regression function, see e.g. Heuchenne and Van Keilegom (2007), or condition (2.2) in Brunel and Comte (2006) which can be interpreted as such kind of truncation. Until now, the truncation bounds which were used were arbitrary fixed, and usually no method is proposed to discuss a method for choosing this truncation bound in practical situations. Therefore, in the new method we propose, we also provide a data-driven procedure to choose the truncation parameter. In our practical implementations, we used a criterion based on an asymptotic discussion which focuses on the mean-squared error associated with the estimation of the single index parameter. We also suggest some possible adaptations to other type of criterion which are covered by our theoretical results.

In section 2, we introduce our censored regression model and present our estimation procedure. It relies on the Kaplan-Meier estimator (1958) of the distribution function, and on semiparametric estimators of the conditional density. Following the procedure of Delecroix *et al.* (2003), we considered kernel based estimators. Our theoretical results are presented in section 3. In section 4 we report simulation results and analysis on real data. Section 5 contains the detailed proof of our Main Lemma which states the asymptotic equivalence of estimating the parameter in the semiparametric and parametric models. All the technicalities are postponed to the section 7.

## 2 Censored regression model and estimation procedure

### 2.1 Notations and general setting

Let  $Y_1, \dots, Y_n$  be i.i.d. copies of a random response variable  $Y \in \mathbb{R}$ , and let  $X_1, \dots, X_n$  be i.i.d. copies of a random vector of covariates  $X \in \mathcal{X}$ , where  $\mathcal{X}$  is a compact subset of  $\mathbb{R}^d$ . Introducing  $C_1, \dots, C_n$  i.i.d. replications of the censoring variable  $C \in \mathbb{R}$ , we consider the following censored

regression framework, where the observations are

$$\begin{cases} Z_i = Y_i \wedge C_i & 1 \leq i \leq n, \\ \delta_i = \mathbb{1}_{\{Y_i \leq C_i\}} & 1 \leq i \leq n, \\ X_i \in \mathcal{X} \subset \mathbb{R}^d & 1 \leq i \leq n. \end{cases}$$

Let us introduce some notations for the distribution functions of the random variables appearing in this model, that is  $H(t) = \mathbb{P}(Z \leq t)$ ,  $F_X(t) = \mathbb{P}(X \leq t)$ ,  $F_Y(t) = \mathbb{P}(Y \leq t)$ ,  $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$  and  $G(t) = \mathbb{P}(C \leq t)$ . A major difficulty arising in censored regression models stands in the unavailability of the empirical distribution function to estimate functions  $F_Y$ ,  $F_{X,Y}$  and  $G$ , which must be replaced by Kaplan-Meier estimators.

We are interested in estimating  $f(y|x)$ , where  $f(y|x)$  denotes the conditional density of  $Y$  given  $X = x$  evaluated at point  $y$ . If one has no insight on the function  $f$ , it becomes necessary to perform nonparametric estimation of the conditional density. In absence of censoring, a classical way to proceed is to use kernel smoothing, see e.g. Bashtannyk and Hyndman (2001). However, the so-called “curse of dimensionality” prevents this approach from being of practical interest when the number of covariates is important ( $d > 3$  in practice). Therefore it becomes relevant to consider semiparametric models which appear to be a good compromise between the parametric (which relies on strong assumptions on the function  $f$  which may not hold in practice) and the nonparametric approach (which relies on fewer assumptions). In the following, we will consider the following semiparametric single-index regression model,

$$\exists \theta_0 \in \Theta \subset \mathbb{R}^d \text{ s.a. } f(y|x) = f_{\theta_0}(y, x'\theta_0), \quad (1)$$

where  $f_{\theta}(y, u)$  denotes the conditional density of  $Y$  given  $X'\theta = u$  evaluated at  $y$ . For identifiability reasons, we will impose that the first component of  $\theta_0$  is one. In comparison to Cox regression model for absolute continuous variables, our model (1) is more general, since it only assumes that the law of  $Y$  given  $X$  depends on an unknown linear combination of the covariates, without imposing additional conditions on the conditional hazard rate.

Model (1) has been considered by Delecroix *et al.* (2003) in the uncensored case. However, their procedure can not be directly applied in the censored framework since the responses variables are not directly observed. As a consequence, the empirical distribution function is unavailable, and most of the tools used in this context are not at our disposal. A solution consists of using procedures relying on Kaplan-Meier estimators for the distribution function. An important

difficulty arising in this type of techniques stands in the poor behavior of Kaplan-Meier estimators in the tail of the distribution. A practical way to prevent us from this kind of drawback is to consider truncated version of the variable  $Y$ . In the following, we will consider  $A_\tau$  a sequence of compacts included in the set  $\{t : \tau_1 \leq t \leq \tau\}$ , for  $\tau \leq \tau_0$ , where  $\tau_0 < \inf\{t : H(t) = 1\}$ . Using only the observations in  $A_\tau$  allow us to avoid the bad behavior of usual Kaplan-Meier estimators in the tail of the distribution. Moreover, this technique of truncation is particularly adapted to our problem of estimating  $\theta_0$ . In our framework, this truncation does not lead to any asymptotic bias, since, denoting by  $f^\tau(\cdot|x)$  the conditional density of  $Y$  given  $X = x$  and  $Y \in A_\tau$ , for any  $\tau < \infty$ , we have, under (1),

$$f^\tau(y|x) = f_{\theta_0}^\tau(y, x'\theta_0), \quad (2)$$

where  $f_\theta^\tau(y, u)$  denotes the conditional density of  $Y$  given  $X'\theta = u$  and  $Y \in A_\tau$  evaluated at  $y$ , and where the parameter is the same in (1) as in (2). In section 2.6, we will discuss a new method allowing to choose  $\tau$  from the data in order to improve the performance in estimating  $\theta_0$ .

## 2.2 Estimation procedure

We will extend the idea behind the procedure developed by Delecroix *et al.* (2003), adapting it to our censored framework. First assume that we know the family of functions  $f_\theta^\tau$ . This approach is a modification of the maximum likelihood estimation procedure. Define, for any function  $J \geq 0$ ,

$$L^\tau(\theta, J) = E [\log f_\theta^\tau(Y_i, \theta'X_i)J(X_i)\mathbb{1}_{Y_i \in A_\tau}] = \int \log f_\theta^\tau(y, \theta'x)J(x)\mathbb{1}_{y \in A_\tau}dF_{X,Y}(x, y).$$

Here,  $J$  is a positive trimming function which will be defined later in order to avoid denominators problems in the nonparametric part of the model, see section 2.4. From (2),  $\theta_0$  maximizes  $L^\tau(\theta, J)$  for any  $\tau < \infty$ , this maximum being unique under some additional conditions on the regression model and  $J$ . Since, in our framework,  $F_{X,Y}$  and  $f_\theta^\tau$  are unknown, it is natural to estimate them in order to produce an empirical version of  $L^\tau(\theta, J)$ .

### 2.2.1 Estimation of $F_{X,Y}$

In the case where there is no censoring (as in Delecroix *et al.* (2003)),  $F_{X,Y}$  can be estimated by the empirical distribution function. In our censoring framework, the empirical distribution function of  $(X, Y)$  is unavailable, since it relies on the true  $Y_i'$ s which are not observed. A

convenient way to proceed consists of replacing it by some Kaplan-Meier estimator such as the one proposed by Stute (1993). Let us define the Kaplan-Meier estimator (Kaplan and Meier (1958)) of  $F_Y$ ,

$$\begin{aligned}\hat{F}_Y(y) &= 1 - \prod_{i: Z_i \leq y} \left( 1 - \frac{1}{\sum_{j=1}^n \mathbb{1}_{Z_j \geq Z_i}} \right)^{\delta_i} \\ &= \sum_{i=1}^n \delta_i W_{in} \mathbb{1}_{Z_i \leq y},\end{aligned}$$

where  $W_{in}$  denotes the “jump” of Kaplan-Meier estimator at observation  $i$  (see Stute (1993)).

To estimate  $F_{X,Y}$ , Stute proposes to use

$$\hat{F}(x, y) = \sum_{i=1}^n \delta_i W_{in} \mathbb{1}_{Z_i \leq y, X_i \leq x}.$$

Let us also define the following (uncomputable) estimator of the distribution function,

$$\tilde{F}(x, y) = \sum_{i=1}^n \delta_i W_i^* \mathbb{1}_{Z_i \leq y, X_i \leq x},$$

where  $W_i^* = n^{-1}[1 - G(Z_i -)]^{-1}$ . The link between  $\hat{F}$  and  $\tilde{F}$  comes from the fact that, in the case where  $\mathbb{P}(Y = C) = 0$ ,

$$W_{in} = n^{-1}[1 - \hat{G}(Z_i -)]^{-1}, \quad (3)$$

where  $\hat{G}$  denotes the Kaplan-Meier estimator of  $G$  (see Satten and Datta (2001)). Asymptotic properties of  $\hat{F}$  can be deduced from studying the difference with the simplest but uncomputable estimator  $\tilde{F}$ .

If we know the family of regression functions  $f_\theta^\tau$ , it is possible to compute the empirical version of  $L^\tau(\theta, J)$  using  $\hat{F}$ , that is

$$\begin{aligned}L_n^\tau(\theta, f^\tau, J) &= \int \log f_\theta^\tau(y, \theta'x) J(x) \mathbb{1}_{y \in A_\tau} d\hat{F}(x, y) \\ &= \sum_{i=1}^n \delta_i W_{in} \log f_\theta^\tau(Z_i, \theta'X_i) J(X_i) \mathbb{1}_{Z_i \in A_\tau}.\end{aligned}$$

In the case  $J \equiv 1$ , the estimator of  $\theta_0$  obtained by maximizing  $L_n^\tau$  would turn out to be an extension of the maximum likelihood estimator of  $\theta_0$ , used in presence of censoring.

### 2.3 Estimation of $f_\theta^\tau$

In our regression model (2), the family  $\{f_\theta^\tau, \theta \in \Theta\}$  is actually unknown. As in Delecroix *et al.* (2003), we propose to use nonparametric kernel smoothing to estimate  $f_\theta^\tau$ . Introducing a kernel

function  $K$  and a sequence of bandwidths  $h$ , define

$$\hat{f}_\theta^{h,\tau}(z, \theta'x) = \frac{\int K_h(\theta'x - \theta'u)K_h(z - y)\mathbb{1}_{y \in A_\tau}d\hat{F}(u, y)}{\int K_h(\theta'x - \theta'u)\mathbb{1}_{y \in A_\tau}d\hat{F}(u, y)}, \quad (4)$$

where  $K_h(\cdot) = h^{-1}K(\cdot/h)$ . Also define  $f^{*h,\tau}$  the kernel estimator based on function  $\tilde{F}$ , that is

$$f_\theta^{*h,\tau}(z, \theta'x) = \frac{\int K_h(\theta'x - \theta'u)K_h(z - y)\mathbb{1}_{y \in A_\tau}d\tilde{F}(u, y)}{\int K_h(\theta'x - \theta'u)\mathbb{1}_{y \in A_\tau}d\tilde{F}(u, y)}.$$

$f^{*h,\tau}$  will play an important role in studying the asymptotic behavior of  $\hat{f}^{h,\tau}$ . Indeed,  $f^{*h,\tau}$  is theoretically more easy to handle with, since it relies on sums of i.i.d. quantities, which is not the case for  $\hat{F}$ . Since  $f^{*h,\tau}$  can be studied by standard kernel arguments, the most important difficulty will arise from studying the difference between  $\hat{f}^{h,\tau}$  and  $f^{*h,\tau}$ .

In the following, we will impose the conditions below on the kernel function.

**Assumption 1.** *Assume that*

(A1)  *$K$  is a twice differentiable and four order kernel with derivatives of order 0, 1 and 2 of bounded variation. Its support is contained in  $[-1/2, 1/2]$  and  $\int_{\mathbb{R}} K(s)ds = 1$ ,*

(A2)  $\|K\|_\infty := \sup_{x \in \mathbb{R}} |K(x)| < \infty$ ,

(A3)  $\mathcal{K} := \{K((x - \cdot)/h) : h > 0, x \in \mathbb{R}^d\}$  is a pointwise measurable class of functions,

(A4)  $h \in \mathcal{H}_n \subset [an^{-\alpha}, bn^{-\alpha}]$  with  $a, b \in \mathbb{R}$ ,  $1/8 < \alpha < 1/6$  and where  $\mathcal{H}_n$  is of cardinality  $k_n$  satisfying  $k_n n^{-4\alpha} \rightarrow 0$ .

## 2.4 The trimming function $J$

The reason behind introducing function  $J$  has to be connected with the need to prevent us from denominators close to zero in the definition (4). Ideally, we would need to use the following trimming function,

$$J_0(x, c) = \tilde{J}(f_{\theta'_0 X}, \theta'_0 x, c), \quad (5)$$

where  $c$  is a strictly positive constant,  $f_{\theta'_0 X}$  denotes the density of  $\theta'_0 X$  and  $\tilde{J}(g, u, c) = \mathbb{1}_{g(u) > c}$ . Unfortunately, this function relies on the knowledge of parameter  $\theta_0$  and  $f_{\theta'_0 X}$ . Therefore, we will have to proceed in two steps, that is first obtain a preliminary consistent estimator of  $\theta_0$ , and then use it to estimate the trimming function  $J_0$  which will be needed to achieve asymptotic normality of our estimators of  $\theta_0$ .

We will assume that we know some set  $B$  on which  $\inf\{f_{\theta'X}(\theta'x) : x \in B, \theta \in \Theta\} > c$ , where  $c$  is a strictly positive constant. In a preliminary step, we can use this set  $B$  to compute the preliminary trimming  $J_B(x) = \mathbb{1}_{x \in B}$ . Using this trimming function, and a deterministic sequence of bandwidth  $h_0$  satisfying (A4) in Assumption 1, we define a preliminary estimator  $\theta_n$  of  $\theta_0$ ,

$$\theta_n = \arg \min_{\theta \in \Theta} L_n^\tau(\theta, \hat{f}^{h_0, \tau}, J_B). \quad (6)$$

Let us stress the fact that  $B$  is assumed to be known by the statistician. This is a classical assumption in single-index regression (see Delecroix *et al.*, (2006)). However, in practice, the procedure does not seem very sensitive to the choice of  $B$ . The bandwidth  $h_0$  we consider in the preliminary step can be any sequence decreasing to zero slower than  $n^{-1/2}$ . Adaptive choice of  $h_0$  could be considered (using, for instance, the same choice as in the final estimation step, see below). However, since we will only need  $\theta_n$  to be a preliminary consistent estimator, and the final estimator will not be very sensitive to an adaptive choice of  $h_0$  while computing  $\theta_n$ , we do not consider this case in the following.

With, at hand, this preliminary estimator  $\theta_n$ , we can compute an estimated version of  $J_0$  which will happen to be equivalent to  $J_0$  (see Delecroix *et al.* (2006) page 738), that is

$$\hat{J}_0(x, c) = \tilde{J}(\hat{f}_{\theta'_n X}^{h_0, \tau}, \theta'_n x, c). \quad (7)$$

For each sequence of bandwidths satisfying (A4) in Assumption 1, and for each truncation bound  $\tau$ , we can define an estimator of  $\theta_0$

$$\hat{\theta}^\tau(h) = \arg \max_{\theta \in \Theta_n} L_n^\tau(\theta, \hat{f}^{h, \tau}, \hat{J}_0), \quad (8)$$

where  $\Theta_n$  is a shrinking sequence of neighborhoods accordingly to the preliminary estimation. However, as for any smoothing approach, the performance of this procedure strongly depends on the bandwidth sequence. Therefore it becomes particularly relevant to provide an approach which automatically selects the most adapted bandwidth according to the data. Then, the new question arising from the censored framework comes from the adaptive choice of the truncation parameter  $\tau$ .

## 2.5 Adaptive choice of the bandwidth

Our procedure consists of choosing from the data, for each  $\theta$ , a bandwidth which is adapted to the computation of  $f_\theta^\tau(z, u)$ . For this, we use an adaptation of the cross-validation technique of



Fan and Yim (2004), that is

$$\hat{h}^\tau(\theta) = \arg \min_{h \in \mathcal{H}_n} \sum_{i=1}^n W_{in} \mathbb{1}_{Z_i \in A_\tau} \left\{ \int_{A_\tau} \hat{f}_\theta^{h,\tau}(z, \theta' X_i)^2 dz - 2 \hat{f}_\theta^{h,\tau}(Z_i, \theta' X_i) \right\}.$$

This criterion is (up to a quantity which does not depend on  $h$ ) an empirical version of the ISE criterion defined in (3.3) in Fan and Yim (2004) (in a censored framework), that is  $\int_{A_\tau} \int \{\hat{f}_\theta^{h,\tau}(z, \theta' x) - f_\theta^\tau(z, \theta' x)\}^2 f_{\theta'X}(\theta' x) dx dz$ .

The estimator of  $\theta_0$  with an adaptive bandwidth is now defined as

$$\hat{\theta}^\tau = \arg \max_{\theta \in \Theta_n} L_n^\tau(\theta, \hat{f}^{h,\tau}, \hat{J}_0). \quad (9)$$

In the above notation,  $\hat{h}$  depends on  $\theta$  and  $\tau$ , which was not emphasized to shorten the notation.

## 2.6 Adaptive choice of $\tau$

As we already mentioned, the Kaplan-Meier estimator does not behave well in the tail of the distribution. For example, if some moment conditions are not satisfied, it is not even  $n^{1/2}$ -consistent. Moreover, even in the case where an appropriate moment condition holds, it may happen (at least for a finite sample size) that the weights corresponding to the large observations are too important and considerably influence the estimation procedure. For this reason, we introduced a truncation by a bound  $\tau$ . However, a large number of existing procedure which also rely on such kind of truncation do not consider the problem of choosing  $\tau$  from the data. We propose to select  $\tau$  from the data in the following way. Suppose that we have a consistent estimator of the asymptotic mean squared error,

$$E^2(\tau) = \limsup_n E \left[ \|\hat{\theta}^\tau(\hat{h}^\tau) - \theta_0\|^2 \right],$$

say  $\hat{E}^2(\tau)$  satisfying

$$\sup_{\tau_1 \leq \tau \leq \tau_0} |\hat{E}^2(\tau) - E^2(\tau)| \rightarrow 0, \text{ in probability.} \quad (10)$$

Such an estimator will be proposed in section 4. Using this empirical estimator, we propose to choose  $\tau$  in the following way, that is

$$\hat{\tau} = \arg \min_{\tau_1 \leq \tau \leq \tau_0} \hat{E}^2(\tau).$$

Our final estimator of  $\theta_0$  is based on an adaptive bandwidth and an adaptive choice of truncation parameter  $\tau$ , that is

$$\hat{\theta} = \hat{\theta}^{\hat{\tau}}.$$

As we already said, truncating the data does not introduce additional bias in the estimation of  $\theta_0$ . On the other hand, removing too many data points could strongly increase the variance and removing some of the largest data points will decrease it. Then, our selection procedure  $\hat{\tau}$  is based on estimating the variance of  $\hat{\theta}$  and consists of taking from the data the truncation parameter  $\tau$  that seems to be the best compromise between these two aspects.

### 3 Asymptotic results

#### 3.1 Consistency

The assumptions needed for consistency can basically be split into three categories, that is identifiability assumptions, assumptions on the regression model (2) itself and finally assumptions on the censoring model.

**Identifiability assumption and assumption on the regression model.**

**Assumption 2.** Assume that for all  $\tau_1 \leq \tau \leq \tau_0$  and all  $\theta \in \Theta - \{\theta_0\}$ ,

$$L_\tau(\theta_0, J_B) - L_\tau(\theta, J_B) > 0.$$

**Assumption 3.** Assume that for  $\theta_1, \theta_2 \in \Theta$ , for a bounded function  $\Phi(X)$  and for some  $\gamma > 0$ , we have

$$\sup_{\tau} \|f_{\theta_1}^\tau(y, \theta_1'x) - f_{\theta_2}^\tau(y, \theta_2'x)\|_\infty \leq \|\theta_1 - \theta_2\|^\gamma \Phi(X).$$

**Assumptions on the censoring model.**

**Assumption 4.**  $\mathbb{P}(Y = C) = 0$ .

This classical assumption in a censored framework avoids problems caused by the lack of symmetry between  $C$  and  $Y$  in the case where there are ties.

**Assumption 5.** *Identifiability assumption : we assume that*

- $Y$  and  $C$  are independent.
- $\mathbb{P}(Y \leq C|X, Y) = \mathbb{P}(Y \leq C|Y)$ .

This last assumption was initially introduced by Stute (1993). An important particular case in which assumption 5 holds is when  $C$  is independent from  $(X, Y)$ . However, assumption 5 is a more general and widely accepted assumption, which allows the censoring variables to depend on the covariates.

**Theorem 1.** *Under Assumptions 2 to 5,*

$$\sup_{\theta \in \Theta, \tau_1 \leq \tau \leq \tau_0} |L_n^\tau(\theta, \hat{f}^{h_0, \tau}, J_B) - L^\tau(\theta, J_B)| = o_P(1), \quad (11)$$

and consequently,

$$\theta_n \rightarrow_{\mathbb{P}} \theta_0.$$

*Proof.* To show (11), we will proceed in two steps. First we consider  $L_n^\tau(\theta, f^\tau, J_B) - L^\tau(\theta, J_B)$  (parametric problem), and then  $L_n^\tau(\theta, \hat{f}^{h_0, \tau}, J_B) - L_n^\tau(\theta, f^\tau, J_B)$ .

**Step 1.** From Assumption 3, the family  $\{\log(f_\theta^\tau(\cdot, \theta' \cdot)), \theta \in \Theta, \tau_1 \leq \tau \leq \tau_0\}$  is seen to be  $P$ -Glivenko-Cantelli. Using an uniform version of Stute (1993) leads to  $\sup_\theta |L_n^\tau(\theta, f^\tau, J_B) - L^\tau(\theta, J_B)| \rightarrow_{\mathbb{P}} 0$ .

**Step 2.** We have, on the set  $\Theta' B$ ,

$$|\log \hat{f}_\theta^{h_0, \tau}(y, u) - \log f_\theta^\tau(y, u)| \leq c^{-1} [\hat{f}_\theta^{h_0, \tau}(y, u) - f_\theta^\tau(y, u)].$$

Hence,

$$\begin{aligned} \sup_{\theta, \tau} |L_n^\tau(\theta, \hat{f}^{h_0, \tau}, J_B) - L_n^\tau(\theta, f^\tau, J_B)| &\leq c^{-1} \sup_{\theta, y, u, \tau} |\hat{f}_\theta^{h_0, \tau}(y, u) - f_\theta^\tau(y, u)| \mathbb{1}_{u \in \Theta' B, y \leq \tau} \int d\hat{F}(x, y) \\ &\leq c^{-1} \sup_{\theta, y, u, \tau} |\hat{f}_\theta^{h_0, \tau}(y, u) - f_\theta^\tau(y, u)| \mathbb{1}_{u \in \Theta' B, y \leq \tau}. \end{aligned}$$

Using the uniform convergence of  $\hat{f}_\theta^{h, \tau}$  (see Proposition 6 and Lemma 7), deduce that

$$\sup_{\theta, \tau} |L_n^\tau(\theta, \hat{f}^{h_0, \tau}, J_B) - L_n^\tau(\theta, f^\tau, J_B)| \rightarrow_{\mathbb{P}} 0. \quad \square$$

### 3.2 Asymptotic normality

To obtain the asymptotic normality of our estimator, we need to add some regularity assumptions on the regression model.

**Assumption 6.** Denote by  $\nabla_\theta f_\theta^\tau(y, x)$  (resp.  $\nabla_\theta^2 f_\theta^\tau(y, x)$ ) the vector of partial derivatives (resp. the matrix of second derivatives with respect to  $\theta$ ) of  $f_\theta^\tau$  with respect to  $\theta$  and computed at point  $(\theta, x, y)$ . Assume that for  $\theta_1, \theta_2 \in \Theta$ , for a bounded function  $\Phi(X)$  and for some  $\gamma > 0$ , we have

$$\sup_\tau \|\nabla_{\theta_1}^2 f_{\theta_1}^\tau(y, x) - \nabla_{\theta_2}^2 f_{\theta_2}^\tau(y, x)\|_\infty \leq \|\theta_1 - \theta_2\|^\gamma \Phi(X).$$

**Assumption 7.** Using the notation of Van der Vaart and Wellner (1996) in section 2.7, define

$$\begin{aligned} \mathcal{H}_1 &= \mathcal{C}^{1+\delta}(\theta'_0 \mathcal{X} \times A_\tau, M), \\ \mathcal{H}_2 &= x \mathcal{C}^{1+\delta}(\theta'_0 \mathcal{X} \times A_\tau, M) + \mathcal{C}^{1+\delta}(\theta'_0 \mathcal{X} \times A_\tau, M) \end{aligned}$$

Assume that  $f_{\theta_0}^\tau(\cdot, \cdot) \in \mathcal{H}_1$  (as a function of  $\theta'_0 x$  and  $y$ ) and  $\nabla_\theta f_{\theta_0}^\tau(\cdot, \cdot) \in \mathcal{H}_2$ .

If the family of functions  $f^\tau$  was known (parametric problem), the asymptotic normality of  $\hat{\theta}$  could be deduced from elementary results on Kaplan-Meier integrals (see section 7 for some brief review of these results), as in Stute (1999) or in Delecroix *et al.* (2008). Using this kind of results, we can derive the following Lemma (see section 7 for the proof) which is sufficient to obtain the asymptotic law of  $\hat{\theta}$  in the parametric case, from Theorem 1 and 2 of Sherman (1994).

**Lemma 2.** *Under Assumptions 6 and 7, we have the following representations:*

1. on  $o_P(1)$ -neighborhoods of  $\theta_0$ ,

$$L_n^\tau(\theta, f^\tau, J_0) = L^\tau(\theta, J_0) + (\theta - \theta_0)'T_{1n}(\theta) + (\theta - \theta_0)'T_{2n}(\theta)(\theta - \theta_0) + T_{3n}(\theta) + T_{4n}(\theta_0),$$

with  $\sup_{\theta, \tau} |T_{1n}| = O_P(n^{-1/2})$ ,  $\sup_{\theta, \tau} |T_{2n}| = o_P(1)$ ,  $\sup_{\theta, \tau} |T_{3n}| = O_P(n^{-1})$  and  $T_{4n}(\theta_0) = L_n^\tau(\theta_0, f^\tau, J_0)$ .

2. on  $O_P(n^{-1/2})$ -neighborhoods of  $\theta_0$ ,

$$L_n^\tau(\theta, f^\tau, J_0) = n^{-1/2}(\theta - \theta_0)'W_{n, \tau} - \frac{1}{2}(\theta - \theta_0)'V_\tau(\theta - \theta_0) + T_{4n}(\theta_0) + T_{5n}(\theta),$$

with  $\sup_{\theta, \tau} |T_{5n}| = o_P(n^{-1})$ , and defining  $f_1(x, y) = f_{\theta_0}^{\tau^{-1}}(y, \theta_0'x)J_0(x, c)\nabla_\theta f_{\theta_0}^\tau(y, x)$ ,

$$W_{n, \tau} = \frac{1}{n^{1/2}} \sum_{i=1}^n \psi(Z_i, \delta_i, X_i; f_1 \mathbb{1}_{A_\tau}),$$

$$V_\tau = E \left[ f_{\theta_0}^{\tau^{-2}}(Y, \theta_0'X)J_0(X, c)\nabla_\theta f_{\theta_0}^\tau(Y, X)\nabla_\theta f_{\theta_0}^\tau(Y, X)' \mathbb{1}_{Y \in A_\tau} \right],$$

where  $\psi$  is defined in Theorem 4.

In the following Theorem, we show that the semiparametric estimator proposed in section 2 has the same asymptotic law as in the fully parametric case.

**Theorem 3.** *Define  $\tau^* = \arg \min_\tau E^2(\tau)$ . Under Assumptions 1 to 7, we have the following asymptotic i.i.d. representation,*

$$\hat{\theta} - \theta_0 = -\frac{1}{n^{1/2}}V_{\tau^*}^{-1}W_{n, \tau^*} + o_P(n^{-1/2}), \quad (12)$$

where  $V_\tau$  and  $W_{n, \tau}$  are defined in Lemma 2. As a consequence,

$$n^{1/2}(\hat{\theta} - \theta_0) \implies \mathcal{N}(0, \Sigma_{\tau^*})$$

where  $\Sigma_{\tau^*} = V_{\tau^*}^{-1}\Delta_{\tau^*}(f_1)V_{\tau^*}^{-1}$ ,  $\Delta_{\tau^*}(f_1) = \text{Var}(\psi(Z, \delta, X; f_1 \mathbb{1}_{A_{\tau^*}}))$  and  $f_1$  is defined in Lemma 2.

This Theorem is a consequence of the Main Lemma below. This result shows that, asymptotically speaking, maximizing  $L_n^\tau(\theta, \hat{f}^{h,\tau}, J)$  is equivalent to maximizing  $L_n^\tau(\theta, f^\tau, J)$ .

**Main Lemma 1.** *Under Assumptions 1 to 7,*

$$L_n^\tau(\theta, \hat{f}^{h,\tau}, \hat{J}_0) = L_n^\tau(\theta, f^\tau, J_0) + (\theta - \theta_0)' R_{1n}(\theta, h, \tau) + (\theta - \theta_0)' R_{2n}(\theta, h, \tau)(\theta - \theta_0) + \tilde{L}_n^\tau(\theta_0),$$

where

$$\begin{aligned} \sup_{\theta \in \Theta_n, h \in \mathcal{H}_n, \tau_1 \leq \tau \leq \tau_0} R_{1n}(\theta, h, \tau) &= o_P(n^{-1/2}), \\ \sup_{\theta \in \Theta_n, h \in \mathcal{H}_n, \tau_1 \leq \tau \leq \tau_0} R_{2n}(\theta, h, \tau) &= o_P(1). \end{aligned}$$

and

$$\tilde{L}_n^\tau(\theta_0) = A_{1n}^\tau(\theta_0, \hat{f}^{h,\tau}) - B_{4n}^\tau(\theta_0, \hat{f}^{h,\tau})$$

where  $A_{1n}^\tau(\theta_0, \hat{f}^{h,\tau})$  and  $B_{4n}^\tau(\theta_0, \hat{f}^{h,\tau})$  are defined in the proof of this Lemma.

In view of Theorem 1 and 2 of Sherman (1994), this result will allow us to obtain the rate of convergence of our estimators, and then the asymptotic law is the same law as the asymptotic law in the parametric problem.

*Proof of Theorem 3.* Define

$$\begin{aligned} \Gamma_{0n}(\theta, \tau, h) &= L_n^\tau(\theta, \hat{f}^{h,\tau}, \hat{J}_0), \\ \Gamma_{1n}(\theta, \tau) &= L_n^\tau(\theta, \hat{f}^{\hat{h},\tau}, \hat{J}_0), \\ \Gamma_{2n}(\theta) &= L_n^{\hat{\tau}}(\theta, \hat{f}^{\hat{h},\hat{\tau}}, \hat{J}_0). \end{aligned}$$

We now apply Theorem 1 and 2 in Sherman (1994) to  $\Gamma_{in}$ , for  $i = 0, 1, 2$ . From our Main Lemma and Lemma 2, we deduce, that the representation (11) in Theorem 2 of Sherman (1994) holds for  $i = 0, 1, 2$ , on  $O_P(n^{-1/2})$ - neighborhoods of  $\theta_0$ , with  $W_n$  and  $V$  defined in Lemma 2. The asymptotic representation (12) is a by-product of the proof of Theorem 2 in Sherman (1994) and of the i.i.d. representations of Kaplan-Meier integrals (see Theorem 4).  $\square$

## 4 Simulation study and real data analysis

### 4.1 Practical implementation of the adaptive choice of $\tau$

From the proof of Theorem 3, we have the representation

$$\hat{\theta} - \theta_0 = -\frac{1}{n} \sum_{i=1}^n V_\tau^{-1} \psi(Z_i, \delta_i, X_i; f_1 \mathbb{1}_{A_\tau}) + o_P(n^{-1/2}).$$

As in Stute (1995), the function  $\psi$  of Theorem 4 can be estimated from the data in the following way by

$$\hat{\psi}(Z, \delta, X; \hat{f}_1 \mathbb{1}_{A_\tau}) = \frac{\delta \hat{f}_1(X, Z)}{1 - \hat{G}(Z-)} + \int \frac{\int_y^{\tau_0} \int_{\mathcal{X}} \hat{f}_1(x, t) d\hat{F}(x, t) dM^{\hat{G}}(y)}{1 - \hat{H}(y)},$$

where  $\hat{f}_1$  is our kernel estimator of  $f_1$  and  $\hat{H}$  is the empirical estimator of  $H$ . To consistently estimate  $\Delta(f_1)$ , we use the general technique proposed by Stute (1996), that is

$$\hat{\Delta}_\tau(f_1) = \frac{1}{n} \sum_{i=1}^n \left[ \hat{\psi}(Z_i, \delta_i, X_i; \hat{f}_1) - \frac{1}{n} \sum_{i=1}^n \hat{\psi}(Z_i, \delta_i, X_i; \hat{f}_1) \right]^{\otimes 2}, \quad (13)$$

where  $\otimes 2$  denotes the product of the matrix with its transpose. A consistent estimator of  $V_\tau$  can then be computed as

$$\hat{V}_\tau = \int \hat{f}_{\hat{\theta}}^{h, \tau^{-2}}(y, \hat{\theta}'x) \hat{J}_0(x, c) \nabla_{\theta} \hat{f}_{\hat{\theta}}^{h, \tau}(y, x) \nabla_{\theta} \hat{f}_{\hat{\theta}}^{h, \tau}(y, x)' \mathbb{1}_{y \in A_\tau} d\hat{F}(x, y).$$

To estimate the asymptotic mean squared error we use

$$\hat{E}_\tau^2 = \frac{1}{n} \hat{W}_{n, \tau}' \hat{V}_\tau^{-1} \hat{V}_\tau^{-1} \hat{W}_{n, \tau}.$$

## 4.2 Simulation study

In order to check the finite sample behavior of our estimators of  $\theta_0$ , we conducted some simulations using a similar model as the one in Delecroix *et al.* (2003). We considered the following regression model,

$$Y_i = \theta_0' X_i + \varepsilon_i, \quad i = 1, \dots, n$$

where  $Y_i \in \mathbb{R}$ ,  $\theta_0 = (1, 0.5, 1.4, 0.2)'$  and  $X_i \sim \otimes^4 \{0.2 \mathcal{N}(0, 1) + 0.8 \mathcal{N}(0.25, 2)\}$ . The errors are centered and normally distributed with conditional variance equal to  $|\theta_0' X|$ . We used the kernel

$$K(u) = 2k(u) - k * k(u)$$

where  $*$  denotes the convolution product and

$$k(u) = \frac{3}{4}(1 - u^2) \mathbb{1}_{|u| \leq 1}$$

is the classical Epanechnikov kernel. The censoring distribution was selected to be exponential with parameter  $\lambda$  which allows us to fix the proportion of censored responses ( $p = 25\%$  and  $p = 40\%$  in our simulations).  $\hat{h}$  was chosen using a regular grid between 1 and 1.5.

Our estimator  $\hat{\theta}^{\hat{\tau}}$  was compared with two other estimators, that is  $\hat{\theta}^\infty$  which does not rely on an adaptive choice of  $\tau$ , and  $\hat{\theta}^{ADE}$  which is obtained using the average derivative method

of Lu and Burke (2005). In the tables below we report our results over 100 simulations from samples of size 100 and 200 for two different rates of censoring. Recalling that the first component of  $\theta_0$  is imposed to be one, we only have to estimate the three other components. For each estimator, the Mean Squared Error  $E(\|\hat{\theta} - \theta_0\|^2)$  is decomposed into bias and covariance.

| $p = 25\%, n = 100$         | Bias   | Variance  | MSE       |
|-----------------------------|--|---|-----------|
| $\hat{\theta}^{ADE}$        | $\begin{pmatrix} -0.112 \\ -0.551 \\ -0.155 \end{pmatrix}$ | $\begin{pmatrix} 0.14 & 0.005 & -0.022 \\ 0.005 & 0.075 & 0.016 \\ -0.022 & 0.016 & 0.116 \end{pmatrix}$  | 0.6714181 |
| $\hat{\theta}^\infty$       | $\begin{pmatrix} 0.057 \\ 0.215 \\ 0.048 \end{pmatrix}$    | $\begin{pmatrix} 0.033 & 0.012 & 0.001 \\ 0.012 & 0.073 & -0.004 \\ 0.001 & -0.004 & 0.027 \end{pmatrix}$ | 0.1841227 |
| $\hat{\theta}^{\hat{\tau}}$ | $\begin{pmatrix} 0.07 \\ 0.221 \\ 0.028 \end{pmatrix}$     | $\begin{pmatrix} 0.034 & 0.002 & 0.002 \\ 0.002 & 0.074 & 0 \\ 0.002 & 0 & 0.02 \end{pmatrix}$            | 0.1825980 |

Table 1: Biases, variances and mean squared errors for 25% of censoring and sampling of size 100.

| $p = 40\%, n = 100$         | Bias   | Variance  | MSE       |
|-----------------------------|--|---|-----------|
| $\hat{\theta}^{ADE}$        | $\begin{pmatrix} -0.334 \\ -0.743 \\ -0.158 \end{pmatrix}$ | $\begin{pmatrix} 0.159 & 0.009 & -0.014 \\ 0.009 & 0.268 & 0.048 \\ -0.014 & 0.048 & 0.165 \end{pmatrix}$   | 1.280163  |
| $\hat{\theta}^\infty$       | $\begin{pmatrix} 0.127 \\ 0.296 \\ 0.096 \end{pmatrix}$    | $\begin{pmatrix} 0.11 & -0.034 & -0.01 \\ -0.034 & 0.101 & 0.021 \\ -0.01 & 0.021 & 0.059 \end{pmatrix}$    | 0.3829797 |
| $\hat{\theta}^{\hat{\tau}}$ | $\begin{pmatrix} 0.074 \\ 0.176 \\ 0.061 \end{pmatrix}$    | $\begin{pmatrix} 0.064 & -0.005 & -0.004 \\ -0.005 & 0.051 & 0.014 \\ -0.004 & 0.014 & 0.069 \end{pmatrix}$ | 0.2239023 |

Table 2: Biases, variances and mean squared errors for 40% of censoring and sampling of size 100.

| $p = 25\%, n = 200$         | Bias   | Variance  | MSE       |
|-----------------------------|--|---|-----------|
| $\hat{\theta}^{ADE}$        | $\begin{pmatrix} -0.189 \\ -0.578 \\ -0.133 \end{pmatrix}$ | $\begin{pmatrix} 0.096 & 0.003 & 0.006 \\ 0.003 & 0.148 & -0.016 \\ 0.006 & -0.016 & 0.131 \end{pmatrix}$ | 0.7620268 |
| $\hat{\theta}^{\infty}$     | $\begin{pmatrix} 0.073 \\ 0.133 \\ 0.015 \end{pmatrix}$    | $\begin{pmatrix} 0.033 & 0.004 & -0.004 \\ 0.004 & 0.023 & 0.002 \\ -0.004 & 0.002 & 0.012 \end{pmatrix}$ | 0.0910719 |
| $\hat{\theta}^{\hat{\tau}}$ | $\begin{pmatrix} 0.034 \\ 0.107 \\ 0.014 \end{pmatrix}$    | $\begin{pmatrix} 0.007 & 0.001 & 0.004 \\ 0.001 & 0.011 & 0 \\ 0 & 0 & 0.006 \end{pmatrix}$               | 0.0364064 |

Table 3: Biases, variances and mean squared errors for 25% of censoring and sampling of size 200.

| $p = 40\%, n = 200$         | Bias   | Variance  | MSE        |
|-----------------------------|--|---|------------|
| $\hat{\theta}^{ADE}$        | $\begin{pmatrix} -0.109 \\ -0.763 \\ -0.053 \end{pmatrix}$ | $\begin{pmatrix} 0.146 & -0.02 & 0.056 \\ -0.02 & 0.143 & -0.014 \\ 0.056 & -0.014 & 0.192 \end{pmatrix}$ | 1.078027   |
| $\hat{\theta}^{\infty}$     | $\begin{pmatrix} 0.104 \\ 0.151 \\ 0.077 \end{pmatrix}$    | $\begin{pmatrix} 0.109 & 0.008 & 0.042 \\ 0.008 & 0.049 & 0.003 \\ 0.042 & 0.003 & 0.055 \end{pmatrix}$   | 0.2521227  |
| $\hat{\theta}^{\hat{\tau}}$ | $\begin{pmatrix} 0.043 \\ 0.14 \\ 0.021 \end{pmatrix}$     | $\begin{pmatrix} 0.018 & -0.001 & 0.002 \\ -0.001 & 0.022 & 0.002 \\ 0.002 & 0.002 & 0.014 \end{pmatrix}$ | 0.07533921 |

Table 4: Biases, variances and mean squared errors for 40% of censoring and sampling of size 200.

To give a precise idea of the number of observations which are removed from the study by choosing  $\tau$  adaptively, introduce  $N = \#\{1 \leq i \leq n, Z_i \leq \hat{\tau}\}$ . In the following table 5, we evaluated  $E[N]$  in the different cases we considered in the simulation study. We also mention the average weight allocated to the largest (uncensored) data point, first in the case where we consider the whole data set (we denote it  $\text{Weight}^{\infty}$ ), then in the truncated data set where we



removed all data points with  $Z_i \geq \hat{\tau}$  (we denote it  $\text{Weight}^{\hat{\tau}}$ ).

|                     | $\hat{\mathbb{E}}(N)$ | $\text{Weight}^{\infty}$ | $\text{Weight}^{\hat{\tau}}$ |
|---------------------|-----------------------|--------------------------|------------------------------|
| $n = 100, p = 25\%$ | 90                    | 0.0667                   | 0.0204                       |
| $n = 100, p = 40\%$ | 87                    | 0.124                    | 0.0236                       |
| $n = 200, p = 25\%$ | 185                   | 0.0402                   | 0.0119                       |
| $n = 200, p = 40\%$ | 172                   | 0.0997                   | 0.0122                       |

Table 5: Last observed data in the truncating model and weight allocated to the largest observation in each model for different sample sizes and censoring rates.

Clearly the MSE deteriorates when the percentage of censoring increases. According to the simulations,  $\hat{\theta}^{\hat{\tau}}$  and  $\hat{\theta}^{\infty}$  outperform  $\hat{\theta}^{ADE}$ , while, as expected, choosing adaptively  $\tau$  improves the quality of the estimation. This is not obvious in the case where there are only 25% of censoring. However, in the case where the level of censoring is high, estimation of the tail of the distribution by Kaplan-Meier estimators becomes more erratic, and the importance of choosing a proper truncation appears in the significant difference between the MSE of  $\hat{\theta}^{\hat{\tau}}$  and  $\hat{\theta}^{\infty}$ . Moreover, the importance of truncation becomes obvious if we look at table 5. We see that, in the case where there is 40 % of censoring, the weight allocated to the largest data-point if we do not use truncation can be up to 10 times (approximatively) the weight allocated to the largest observation in the truncated data set. The ratio is less important in the case where there is 25 % of censoring, but still consequent (in this case, the ratio is approximatively 3). Therefore, it seems that, considering the whole data set, the weight allocated to the largest observation can have a too strong influence on the estimation procedure, which explains the difference of performance of the estimators with or without truncation.

### 4.3 Example : Stanford Heart Transplant Data

We now illustrate our method using data from the Stanford Heart Transplant program. This data set was initially studied by Miller and Halpern (1982). 184 of 249 patients in this program received a heart transplantation between October 1967 and February 1980. From this data, we considered the survival time as the response variable  $Z$ , age as the first component of  $X$  and the square of age as the second component. Patients alive beyond February 1980 were considered censored. For easier comparison to previous work on this data set, we concentrate our analysis on the 157 patients out of 184 who had complete tissue typing. Among these 157 cases, 55 were

censored.

Several methods of estimation have already been applied to this data set to estimate the following regression model,

$$Z = \alpha + \beta'X + \varepsilon(X), \quad (14)$$

where  $\beta = (\beta_1, \beta_2)'$ ,  $E[\varepsilon(X)|X] = 0$ , see Miller and Halpern (1982), Wei *et al.* (1990), Stute *et al.* (2000). Furthermore, nonparametric lack-of-fit tests have shown that the regression model (14) seemed reasonable, see Stute *et al.* (2000) and Lopez and Patilea (2008). Therefore it seems to us appropriate to experiment our model on this data set. This strengthens the assumption on the residual, by assuming that  $\varepsilon(X) = \varepsilon(\theta_0'X)$ , where  $\theta_0 = (1, \beta_2/\beta_1)'$ , but allows more flexibility on the regression function.

In the following table, we present our estimators and recall the values of the estimators of  $\beta_2/\beta_1$  for the linear regression model (14). We first computed  $\hat{\theta}^\infty$ , which is our estimator using the whole data set, that is with  $\tau = +\infty$ , and compared it to the one obtained by choosing  $\tau$  from the data as in section 4.1. In this last case,  $\hat{\tau} = Z_{(90)}$  where  $Z_{(i)}$  denotes the  $i$ -th order statistic, this means that it conducted us to remove the 67 largest observations to estimate  $\theta_0$  (but not to estimate Kaplan-Meier weights, which were computed using the whole data set). We computed  $\text{Weight}^\infty = 0.0397$ , and  $\text{Weight}^{\hat{\tau}} = 0.0076$  for the truncated data. Adaptive bandwidth was 1.7 for  $\hat{\theta}^\infty$ , and 1.3 for  $\hat{\theta}^{\hat{\tau}}$ . The estimated value of the mean-squared error was  $E_\infty^2 = 0.1089375$  and  $E_{\hat{\tau}}^2 = 0.01212701$  for  $\hat{\theta}^\infty$  and  $\hat{\theta}^{\hat{\tau}}$  respectively.

|   | Estimator of $\theta_{0,2} = \beta_2/\beta_1$ |
|---|---|
| Miller and Halpern  | -0.01588785                                   |
| Wei <i>et al.</i>   | 63.75   |
| Stute <i>et al.</i>   | -0.01367034                                   |
| $\hat{\theta}^\infty$ (without adaptive choice of $\tau$ )    | -0.07351351                                   |
| $\hat{\theta}^{\hat{\tau}}$ (with adaptive choice of $\tau$ ) | -0.0421508                                    |

Table 6: Comparison of different estimators of  $\theta_{0,2}$ .

Our estimators seem relatively close to the ones obtained by Miller and Halpern (1982) and Stute *et al.* (2000) using respectively the Buckley-James method and the Kaplan-Meier integrals method for the linear regression model.

## 5 Proof of Main Lemma

First, the same arguments as in Delecroix *et al.* (2006) apply to replace  $\hat{J}_0$  by  $J_0$ . Define  $J_\theta(x, c) = \mathbb{1}_{f_{\theta'X}(\theta'x) \geq c}$ . From Assumption 3 on the density of  $\theta'x$ , deduce that, on shrinking neighbourhoods of  $\theta_0$ ,  $J_0(x, c)$  can be replaced by  $J_\theta(x, c/2)$ . Using a Taylor expansion, write

$$\begin{aligned} L_n^\tau(\theta, \hat{f}^{h,\tau}, J_0) - L_n^\tau(\theta, f^\tau, J_0) &= \sum_{i=1}^n \delta_i W_{in} \mathbb{1}_{Z_i \in A_\tau} \log \left( \frac{\hat{f}_\theta^{h,\tau}(Z_i, \theta'X_i)}{f_\theta^\tau(Z_i, \theta'X_i)} \right) J_0(X_i, c) \\ &= \sum_{i=1}^n \frac{\delta_i W_{in} \mathbb{1}_{Z_i \in A_\tau} \left( \hat{f}_\theta^{h,\tau}(Z_i, \theta'X_i) - f_\theta^\tau(Z_i, \theta'X_i) \right) J_0(X_i, c)}{f_\theta^\tau(Z_i, \theta'X_i)} \\ &\quad - \sum_{i=1}^n \frac{\delta_i W_{in} \mathbb{1}_{Z_i \in A_\tau} \left[ \hat{f}_\theta^{h,\tau}(Z_i, \theta'X_i) - f_\theta^\tau(Z_i, \theta'X_i) \right]^2 J_0(X_i, c)}{\phi(f_\theta^\tau(Z_i, \theta'X_i), \hat{f}_\theta^{h,\tau}(Z_i, \theta'X_i))^2} \\ &= A_{1n}^\tau(\theta, \hat{f}^{h,\tau}) - B_{1n}^\tau(\theta, \hat{f}^{h,\tau}) \end{aligned}$$

where  $\phi(f_\theta^\tau(Z_i, \theta'X_i), \hat{f}_\theta^{h,\tau}(Z_i, \theta'X_i))$  is between  $\hat{f}_\theta^{h,\tau}(Z_i, \theta'X_i)$  and  $f_\theta^\tau(Z_i, \theta'X_i)$ .

**Step 1.** We first study  $A_{1n}$ . A Taylor expansion leads to the following decomposition,

$$\begin{aligned} A_{1n}^\tau &= (\theta - \theta_0)' \sum_{i=1}^n \frac{\delta_i W_{in} \mathbb{1}_{Z_i \in A_\tau} (\nabla_\theta \hat{f}_{\theta_0}^{h,\tau}(Z_i, X_i) - \nabla_\theta f_{\theta_0}^\tau(Z_i, X_i)) J_\theta(X_i, c/2)}{f_\theta^\tau(Z_i, \theta'X_i)} \\ &\quad + (\theta - \theta_0)' \left[ \sum_{i=1}^n \frac{\delta_i W_{in} \mathbb{1}_{Z_i \in A_\tau} (\nabla_\theta^2 \hat{f}_{\theta_0}^{h,\tau}(Z_i, X_i) - \nabla_\theta^2 f_{\theta_0}^\tau(Z_i, X_i)) J_\theta(X_i, c/2)}{2f_\theta^\tau(Z_i, \theta'X_i)} \right] (\theta - \theta_0) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \frac{\delta_i W_{in} \mathbb{1}_{Z_i \in A_\tau} (\hat{f}_{\theta_0}^{h,\tau}(Z_i, \theta'_0 X_i) - f_{\theta_0}^\tau(Z_i, \theta'_0 X_i))}{f_\theta^\tau(Z_i, \theta'X_i) f_{\theta_0}^\tau(Z_i, \theta'_0 X_i)} \\ &\quad \times (f_{\theta_0}^\tau(Z_i, \theta'_0 X_i) - f_\theta^\tau(Z_i, \theta'X_i)) J_0(X_i, c) J_\theta(X_i, c/2) + A_{1n}^\tau(\theta_0, \hat{f}^{h,\tau}) \\ &= A_{1n}^\tau(\theta_0, \hat{f}^{h,\tau}) + (\theta - \theta_0)' A_{2n}^\tau(\theta_0, \hat{f}^{h,\tau}) + (\theta - \theta_0)' A_{3n}^\tau(\tilde{\theta}, \hat{f}^{h,\tau})(\theta - \theta_0) + A_{4n}^\tau(\theta, \hat{f}^{h,\tau}), \end{aligned}$$

for some  $\tilde{\theta}$  between  $\theta$  and  $\theta_0$ . Observe that, using the uniform consistency of  $\nabla_\theta^2 \hat{f}_\theta^{h,\tau}$  (deduced from Proposition 6 and Lemma 7), we obtain  $\sup_{\tilde{\theta} \in \Theta_n, \tau \leq \tau_0, h \in \mathcal{H}_n} A_{3n}^\tau(\tilde{\theta}, \hat{f}^{h,\tau}) = o_P(1)$ . We now study  $A_{2n}^\tau(\theta_0, \hat{f}^{h,\tau})$ . Using the expression (3) of the jumps of Kaplan-Meier estimator, observe that

$$\begin{aligned} A_{2n}^\tau(\theta, \hat{f}^{h,\tau}) &= \sum_{i=1}^n \frac{W_i^* \mathbb{1}_{Z_i \in A_\tau} (\nabla_\theta \hat{f}_{\theta_0}^{h,\tau}(Z_i, X_i) - \nabla_\theta f_{\theta_0}^\tau(Z_i, X_i)) J_\theta(X_i, c/2)}{f_\theta^\tau(Z_i, \theta'X_i)} \\ &\quad + \frac{1}{n} \sum_{i=1}^n W_i^* Z_{G(Z_i-)} \frac{\delta_i \mathbb{1}_{Z_i \in A_\tau} (\nabla_\theta \hat{f}_{\theta_0}^{h,\tau}(Z_i, X_i) - \nabla_\theta f_{\theta_0}^\tau(Z_i, X_i)) J_\theta(X_i, c/2)}{f_\theta^\tau(Z_i, \theta'X_i)} \\ &= A_{21n}^\tau(\theta, \hat{f}^{h,\tau}) + A_{22n}^\tau(\theta, \hat{f}^{h,\tau}), \end{aligned}$$

where

$$Z_G(t) = \frac{\hat{G}(t) - G(t)}{1 - \hat{G}(t)}.$$

The term  $A_{22n}^\tau$  can be bounded using (21), (22) and Lemma 7, by

$$\sup_{\tau \leq \tau_0} |A_{22n}(\theta, \hat{f}^{h,\tau})| \leq o_P(n^{-1/2}) \times n^{-1} \sum_{i=1}^n \delta_i [1 - G(Z_i-)]^{-1},$$

and the last term is  $O_P(1)$  since it has finite expectation. Now for  $A_{21n}^\tau$ , first replace  $\theta$  at the denominator by  $\theta_0$ . We have

$$\begin{aligned} A_{21n}^\tau(\theta, \hat{f}^{h,\tau}) &= \sum_{i=1}^n \frac{W_i^* \mathbb{1}_{Z_i \in A_\tau} (\nabla_\theta \hat{f}_{\theta_0}^{h,\tau}(Z_i, X_i) - \nabla_\theta f_{\theta_0}^\tau(Z_i, X_i)) J_0(X_i, c/4)}{f_{\theta_0}^\tau(Z_i, \theta'_0 X_i)} \\ &\quad + R_n^\tau(\theta, h)(\theta - \theta_0), \end{aligned}$$

with  $\sup_{\theta \in \Theta_n, \tau \leq \tau_0, h \in \mathcal{H}_n} |R_n^\tau(\theta, h)| = o_P(1)$  from Assumption 3 and the uniform consistency of  $\nabla_\theta \hat{f}_{\theta_0}^{h,\tau}$  deduced from Proposition 6 and Lemma 7. Then use Assumption 7 and Proposition 9. Using the equicontinuity property of Donsker classes (see e.g. Van der Vaart and Wellner (1996) or Van der Vaart (1998)), we obtain that

$$\begin{aligned} A_{2n}^\tau(\theta, \hat{f}^{h,\tau}) &= \iint \frac{[\nabla_\theta \hat{f}_{\theta_0}^{h,\tau}(y, x) - \nabla_\theta f_{\theta_0}^\tau(y, x)] \mathbb{1}_{y \in A_\tau} J_0(x, c/4) d\mathbb{P}(x, y)}{f_{\theta_0}^\tau(y, u)} \\ &\quad + o_P(n^{-1/2}), \end{aligned}$$

where the  $o_P$ -rate does not depend on  $\theta$ ,  $h$ , nor  $\tau$ . From classical kernel arguments,  $\sup_{y,x,\tau} |\int (\nabla_\theta \hat{f}_{\theta_0}^{h,\tau}(y, x) - \nabla_\theta f_{\theta_0}^\tau(y, x)) \mathbb{1}_{y \in A_\tau} J_0(x, c/4) d\mathbb{P}(x, y)| = O_{\mathbb{P}}(h^4) = o_{\mathbb{P}}(n^{-1/2})$ , since  $nh^8 \rightarrow 0$ . Then, Lemma 8 concludes the proof for  $A_{2n}^\tau(\theta, \hat{f}^{h,\tau})$ .  $A_{4n}^\tau(\theta, \hat{f}^{h,\tau})$  can be handled similarly.

**Step 2.**  $B_{1n}^\tau$  can be rewritten as

$$\begin{aligned} B_{1n}^\tau(\theta, \hat{f}^{h,\tau}) &= \sum_{i=1}^n \delta_i W_{in} \mathbb{1}_{Z_i \in A_\tau} J_\theta(X_i, c/2) \frac{\{(\theta - \theta_0)' [\nabla_\theta \hat{f}_{\theta_0}^{h,\tau}(Z_i, X_i) - \nabla_\theta f_{\theta_0}^\tau(Z_i, X_i)]\}^2}{\phi(f_\theta^\tau(Z_i, \theta' X_i), \hat{f}_\theta^{h,\tau}(Z_i, \theta' X_i))^2} \\ &\quad + 2 \sum_{i=1}^n \delta_i W_{in} J_\theta(X_i, c/2) \mathbb{1}_{Z_i \in A_\tau} [\hat{f}_{\theta_0}^{h,\tau}(Z_i, \theta'_0 X_i) - f_{\theta_0}^\tau(Z_i, \theta'_0 X_i)] \\ &\quad \times (\theta - \theta_0)' [\nabla_\theta \hat{f}_\theta^{h,\tau}(Z_i, X_i) - \nabla_\theta f_\theta^\tau(Z_i, X_i)] [\phi(f_\theta^\tau(Z_i, \theta' X_i), \hat{f}_\theta^{h,\tau}(Z_i, \theta' X_i))^2]^{-1} \\ &\quad + B_{4n}^\tau(\theta_0, \hat{f}^{h,\tau}) + o_P(\|\theta - \theta_0\|^2) \\ &= (\theta - \theta_0)' B_{2n}^\tau(\theta_0, \hat{f}^{h,\tau})(\theta - \theta_0) + (\theta - \theta_0)' B_{3n}^\tau(\theta, \hat{f}^{h,\tau}) + B_{4n}^\tau(\theta_0, \hat{f}^{h,\tau}) + o_P(\|\theta - \theta_0\|^2) \end{aligned}$$

for some  $\tilde{\theta}$  between  $\theta$  and  $\theta_0$ . The third term does not depend on  $\theta$ . For  $B_{2n}^\tau$ , use the uniform consistency of  $\nabla_{\theta_0} \hat{f}_{\theta_0}^{h,\tau}$  (Proposition 6 and Lemma 7) to obtain  $\sup_{\tau \leq \tau_0, h \in \mathcal{H}_n} |B_{2n}^\tau(\theta, \hat{f}^{h,\tau})| = o_P(n^{-1/2})$ . Finally, for  $B_{3n}^\tau(\theta, \hat{f}^{h,\tau})$ , from a Taylor expansion,

$$\begin{aligned} B_{3n}^\tau(\theta, \hat{f}^{h,\tau}) &= 2 \sum_{i=1}^n \frac{\delta_i W_{in} \mathbf{1}_{Z_i \in A_\tau} J_\theta(X_i, c/2)}{\phi(f_\theta^\tau(Z_i, \theta' X_i), \hat{f}_\theta^{h,\tau}(Z_i, \theta' X_i))^2} \\ &\quad \times [\hat{f}_{\theta_0}^{h,\tau}(Z_i, \theta'_0 X_i) - f_{\theta_0}^\tau(Z_i, \theta'_0 X_i)] [\nabla_\theta \hat{f}_{\theta_0}^{h,\tau}(Z_i, X_i) - \nabla_\theta f_{\theta_0}^\tau(Z_i, X_i)] \\ &\quad + (\theta - \theta_0)' R_n^\tau(\theta, \hat{f}^{h,\tau}), \end{aligned}$$

with  $\sup_{\theta \in \Theta_n, \tau \leq \tau_0, h \in \mathcal{H}_n} R_n^\tau(\theta, \hat{f}^{h,\tau}) = o_P(1)$ , from Proposition 6 and Lemma 7. For the main term, the product of the uniform convergence rates of  $\hat{f}_{\theta_0}^{h,\tau}$  and  $\nabla_\theta \hat{f}_{\theta_0}^{h,\tau}$  obtained from Proposition 6 and Lemma 7 is  $o_P(n^{-1/2})$  for  $h \in \mathcal{H}_n$ .

## 6 Conclusion

We proposed a new estimation procedure of a conditional density under a single-index assumption and random censoring. This procedure is an extension of the approach of Delecroix *et al.* (2003) in the case of a censored response. One of the advantage of this model is that it relies on fewer assumptions as a Cox regression model, in the case where the random variables of the model are absolutely continuous. By showing that estimating in this semiparametric model is asymptotically equivalent to estimating in a parametric one (unknown in practice), we obtain a  $n^{-1/2}$ -rate for the estimator of the index. This estimator can then be used to estimate the conditional density or the conditional distribution function by using traditional nonparametric estimator under censoring. A new feature of our procedure, is that it provides an adaptively driven choice of the bandwidth involved in the kernel estimators we used, and that it also provides an adaptive choice of a truncation parameter needed to avoid problems caused by the bad behavior of Kaplan-Meier estimators in the tail of the distribution. In this specific problem, this truncation does not introduce some additional bias in the procedure, and seems, according to our simulations, to increase the quality of the estimator, especially in the case where the proportion of censored responses is important. Our way of choosing  $\tau$  was motivated by minimizing the MSE in the estimation of  $\hat{\theta}$ . However, our method could be easily adapted to other kinds of criteria which, for example more focus on the error in estimating one specific direction, or on the error in the estimation of the conditional density itself.

## 7 Appendix

### 7.1 Kaplan-Meier integrals for the parametric case

We first recall a classical asymptotic representation of integrals with respect to  $\hat{F}$ . See Stute (1995), Stute (1996) and Sánchez Sellero *et al.* (2005).

**Theorem 4.** *Let  $\mathcal{F}$  be a VC-class of functions with envelope  $\Phi$  such as*

$$\Phi(x, y) = 0, \text{ for all } y \geq \tau_0, \quad (15)$$

where  $\tau_0 \leq \tau_H$ . We have the following asymptotic i.i.d. representation, for all  $\phi \in \mathcal{F}$ ,

$$\int \phi(x, y) d\hat{F}(x, y) = \frac{1}{n} \sum_{i=1}^n \psi(Z_i, \delta_i, X_i; \phi) + R(\phi),$$

where  $\sup_{\phi \in \mathcal{F}} |R(\phi)| = O_{a.s.}([\log n]^3 n^{-1})$ , and

$$\psi(Z_i, \delta_i, X_i; \phi) = \frac{\delta \phi(X_i, Z_i)}{1 - G(Z_i -)} + \int \frac{\int_y^{\tau_0} \int_{\mathcal{X}} \phi(x, t) dF(x, t) dM_i^G(y)}{1 - H(y)},$$

where  $M_i^G(y) = (1 - \delta_i) \mathbb{1}_{Z_i \leq y} - \int_{-\infty}^y \mathbb{1}_{Z_i \geq t} [1 - G(t-)]^{-1} dG(t)$  is a martingale with respect to the filtration  $\mathcal{G}_y = \{(Z_i, \delta_i, X_i) \mathbb{1}_{Z_i \leq y}\}$ . Define  $\Delta(\phi) = \text{Var}(\psi(Z, \delta, X; \phi))$ . Then it follows that

$$\sqrt{n} \int \phi(x, y) d[\hat{F} - F](x, y) \implies \mathcal{N}(0, \Delta(\phi)).$$

Initially, the result of Stute was derived for a single function  $\phi$ . Furthermore, Theorem 1.1 in Stute (1996) gives a convergence rate which is only  $o_P(n^{-1/2})$  for the remainder term, however an higher convergence rate is obtained in his proof of Theorem 1.1 for functions satisfying (15), which is the only case considered in our work. To obtain uniformity on a VC-class of functions, see Sánchez Sellero *et al.* (2005) who provided a more general representation that extends the one of Stute in the case where  $Y$  is right-censored and left-truncated. Their result is really useful since it provides, as a corollary, uniform law of large numbers results and uniform central limit theorem. The representation we present in our Theorem 4 is a simple rewriting of Stute's representation. Theorem 4 is then a key ingredient to prove Lemma 2.

*Proof of Lemma 2.* We directly show the second part of the Lemma, since the first can be studied from similar techniques. From a Taylor expansion,

$$\begin{aligned} L_n^\tau(\theta, f^\tau, J_0) &= (\theta - \theta_0)' \sum_{i=1}^n \delta_i W_{in} J_0(X_i, c) \mathbb{1}_{Z_i \in A_\tau} \frac{\nabla_\theta f_{\theta_0}^\tau(Z_i, X_i)}{f_{\theta_0}^\tau(Z_i, \theta_0' X_i)} \\ &\quad + \frac{1}{2} (\theta - \theta_0)' \sum_{i=1}^n \delta_i W_{in} J_0(X_i, c) \mathbb{1}_{Z_i \in A_\tau} \nabla_\theta^2 [\log f_\theta^\tau](Z_i, X_i) (\theta - \theta_0) \\ &\quad + T_{4n}(\theta_0), \end{aligned} \quad (16)$$

for some  $\tilde{\theta}$  between  $\theta_0$  and  $\theta$ . Theorem 4 provides an i.i.d. representation for the first term (which corresponds to  $W_{n,\tau}$  in Lemma 2), while, from Assumption 6, the family of functions  $\nabla_{\theta}^2[\log f_{\theta}^{\tau}](y, x)\mathbb{1}_{y \in A_{\tau}}$  is a  $VC$ -class of functions satisfying (15). Hence the sum in the second term of (16) tends to  $V$  almost surely using an uniform law of large numbers property.  $\square$

## 7.2 The gradient of $f$

In the following for any function  $\varphi$  we will denote by  $\varphi_h^{(n)}(\cdot)$  the expression  $h^{-n}\varphi^{(n)}(\cdot/h)$  such as, for example  $K_h'(\cdot) = h^{-1}K'(\cdot/h)$ .

**Proposition 5.** *Let*

$$f_{\tau}'(y, u) = \partial_u f_{\theta_0}^{\tau}(y, u).$$

*We have*

$$\nabla_{\theta} f_{\theta_0}^{\tau}(y', x) = x f_{1,y,\tau}(y', \theta_0' x) + f_{2,y,\tau}(y', \theta_0' x),$$

*with*

$$f_{1,\tau}(y, \theta_0' x) = f_{\tau}'(y, \theta_0' x),$$

$$f_{2,\tau}(y, \theta_0' x) = -f_{\tau}'(y, \theta_0' x) E[X | \theta_0' X].$$

*In particular,  $E[\nabla_{\theta} f_{\theta_0}^{\tau}(Y, X) | \theta_0' X] = 0$ .*

*Proof.* Direct adaptation of Lemma 5A in Dominitz and Sherman (2005).  $\square$

## 7.3 Convergence properties of $f^{*h,\tau}$

We first recall some classical properties on kernel estimators. Consider the class of functions  $\mathcal{K}$  introduced in Assumption 1. Let  $N(\varepsilon, \mathcal{K}, d_Q)$  be the minimal number of balls  $\{g : d_Q(g, g') < \varepsilon\}$  of  $d_Q$ -radius  $\varepsilon$  needed to cover  $\mathcal{K}$ . For  $\varepsilon > 0$ , let  $N(\varepsilon, \mathcal{K}) = \sup_Q N(\varepsilon, \mathcal{K}, d_Q)$ , where the supremum is taken over all probability measures  $Q$  on  $(\mathbb{R}^d, \mathcal{B})$ ,  $d_Q$  is the  $L_2(Q)$ -metric. From Nolan and Pollard (1987), it can easily be seen that, using a kernel  $K$  satisfying Assumption 1, for some  $C > 0$  and  $\nu > 0$ ,  $N(\varepsilon, \mathcal{K}) \leq C\varepsilon^{-\nu}$ ,  $0 < \varepsilon < 1$ .

**Proposition 6.** *Under assumption 1 we have, for some  $c > 0$*

$$\sup_{x,y,h,\tau} \left| f_{\theta_0}^{*h,\tau}(y, \theta_0' x) - f_{\theta_0}^{\tau}(y, \theta_0' x) \right| \mathbb{1}_{y \in A_{\tau}} J_0(x, c) = O_P \left( n^{-1/2} h^{-1} [\log n]^{1/2} \right), \quad (17)$$

$$\sup_{x,y,h,\tau} \left| \nabla_{\theta} f_{\theta_0}^{*h,\tau}(y, x) - \nabla_{\theta} f_{\theta_0}^{\tau}(y, x) \right| \mathbb{1}_{y \in A_{\tau}} J_0(x, c) = O_P \left( n^{-1/2} h^{-2} [\log n]^{1/2} \right), \quad (18)$$

$$\sup_{x,y,h,\tau,\theta} \left| \nabla_{\theta}^2 f_{\theta_0}^{*h,\tau}(y, x) - \nabla_{\theta}^2 f_{\theta_0}^{\tau}(y, x) \right| \mathbb{1}_{y \in A_{\tau}} J_{\theta}(x, c) = O_P \left( n^{-1/2} h^{-3} [\log n]^{1/2} \right). \quad (19)$$

*Proof.* (17) is a direct application of Theorems 1 and 4 in Einmahl and Mason (2005). For (18), we only show the convergence for the term

$$\hat{r}_{\theta_0}^{h,\tau}(x, y) := \frac{1}{h} \sum_{i=1}^n \delta_i W_i^* \mathbb{1}_{Z_i \in A_\tau} J_0(x, c)(X_i - x) K'_h(X'_i \theta_0 - x' \theta_0) K_h(Z_i - y).$$

Define

$$\bar{r}_{\theta_0}^{h,\tau}(x, y) = \frac{1}{h} \mathbb{E} [\mathbb{1}_{Y \in A_\tau} J_0(x, c)(X - x) K'_h(X' \theta_0 - x' \theta_0) K_h(Y - y)]$$

and

$$r_{\theta_0}^\tau(x, y) = \frac{\partial}{\partial u} \left\{ \mathbb{E} [(X - x) | \theta'_0 X = u, Y = y] \mathbb{1}_{y \in A_\tau} J_0(x, c) f_{\theta'_0 X, Y}(u, y) \right\} \Big|_{u=\theta'_0 x}.$$

Note that, from our assumptions  $r_{\theta_0}^\tau$  is a finite quantity. Next, Theorem 4 in Einmahl and Mason (2005) yields :

$$\sup_{x, y, h, \tau} \left| \hat{r}_{\theta_0}^{h,\tau}(x, y) - \bar{r}_{\theta_0}^{h,\tau}(x, y) \right| \mathbb{1}_{y \in A_\tau} = O_P(n^{-1/2} h^{-2} [\log n]^{1/2}).$$

For the bias term,  $\sup_{x, y, h, \tau} \left| \bar{r}_{\theta_0}^{h,\tau}(x, y) - r_{\theta_0}^\tau(x, y) \right| \mathbb{1}_{y \in A_\tau} = O(h^4) = o(n^{-1/2})$ , (see e.g. Bosq and Lecoutre (1997)). As a consequence,

$$\sup_{x, y, h, \tau} \left| \hat{r}_{\theta_0}^{h,\tau}(x, y) - r_{\theta_0}^\tau(x, y) \right| \mathbb{1}_{y \in A_\tau} = O_P(n^{-1/2} h^{-2} [\log n]^{1/2}).$$

For (19), we also need an uniformity with respect to  $\theta$ . The result can be deduced from the uniform convergence (with respect to  $\theta, x, u$ ) of quantities such as

$$S_n^{h,\tau}(\theta, x, y, \beta) = \frac{1}{h^2} \sum_{i=1}^n \delta_i W_i^* \phi(Z_i, X_i, \theta) \nabla_\theta^\beta K \left( \frac{\theta' X_i - \theta' x}{h} \right) K \left( \frac{Z_i - y}{h} \right), \quad (20)$$

where  $\nabla_\theta^\beta K([\theta' X_i - \theta' x] h^{-1})$  for  $\beta = 1$  (resp. for  $\beta = 2$ ) denotes the gradient vector of function  $K([\theta' X_i - \theta' x]/h)$  (resp. Hessian matrix) with respect to  $\theta$  and evaluated at  $\theta$ , and where  $\phi$  is a bounded function with respect to  $\theta$  and  $x$ . The function  $\phi$  we consider is  $\phi(Z, X, \theta) = f_{\theta' X}^\tau(\theta' X)^{-1} \mathbb{1}_{Z \in A_\tau} J_0(x, c)$  with the convention  $0/0 = 0$  and where  $f_{\theta' X}^\tau(\theta' X)$  is the conditional density of  $\theta' X$  given  $Y \in A_\tau$ . (20) can be studied using the same method as Einmahl and Mason (2005). For this, observe that the family of functions  $\{(X, Z) \rightarrow \nabla_\theta^\beta K([\theta' X - \theta' x] h^{-1}) K([Z - y] h^{-1}), \theta \in \Theta, x, y\}$  satisfies the Assumptions of Proposition 1 in Einmahl and Mason (2005) (see Lemma 22 (ii) in Nolan and Pollard (1987)). Hence, apply Talagrand's inequality (Talagrand (1994), see also Einmahl and Mason (2005)) to obtain that

$$\sup_{\theta, x, y, h, \tau} |S_n^{h,\tau}(\theta, x, y, \alpha) - E[S_n^{h,\tau}(\theta, x, y, \alpha)]| \mathbb{1}_{y \in A_\tau} = O_P(n^{-1/2} [\log n]^{1/2} h^{-1-\beta}).$$

Again, the bias term converges uniformly at rate  $O(h^4)$ . □



## 7.4 The difference between $f^*$ and $\hat{f}$

### 7.4.1 Convergence rate of $\hat{f}$

In this section, we show that replacing  $f^{*h,\tau}$  by  $\hat{f}^{h,\tau}$  (which is the estimator used in practice) does not modify the rate of convergence. To give the intuition of this results, observe that  $f^{*h,\tau}$  was obtained from  $\hat{f}^{h,\tau}$  by replacing  $\hat{G}$  by  $G$ . Let us recall some convergence properties of  $\hat{G}$ . We have

$$\sup_{t \leq \tau_0} |\hat{G}(t) - G(t)| = O_P(n^{-1/2}), \quad (21)$$

$$\sup_{t \leq \tau_0} \frac{1 - G(t)}{1 - \hat{G}(t)} = O_P(1). \quad (22)$$

See Gill (1983) for (21) and Zhou (1992) for (22). From (21), we see that the convergence rate of  $\hat{G}$  is faster than the convergence rate of  $f^{*h,\tau}$ , which explains the asymptotic equivalence of  $\hat{f}^{h,\tau}$  and  $f^{*h,\tau}$ . Lemma 7 makes things more precise and also gives a representation of the difference between  $\nabla_{\theta} f_{\theta_0}^{*h,\tau}$  and  $\nabla_{\theta} \hat{f}_{\theta_0}^{h,\tau}$  which is needed in the proof of Main Lemma. Also required to prove our Main Lemma, Lemma 8 below gives a technical result on the integral of this difference.

**Lemma 7.** *Under the Assumption of Lemma 2, we have for some  $c > 0$*

$$\sup_{x,y,h,\tau} \left| \hat{f}_{\theta_0}^{h,\tau}(y, \theta'x) - f_{\theta_0}^{*h,\tau}(y, \theta'x) \right| \mathbb{1}_{y \in A_{\tau}} J_0(x, c) = O_P(n^{-1/2}), \quad (23)$$

$$\sup_{x,y,h,\tau} \left| \nabla_{\theta} \hat{f}_{\theta_0}^{h,\tau}(y, x) - \nabla_{\theta} f_{\theta_0}^{*h,\tau}(y, x) \right| \mathbb{1}_{y \in A_{\tau}} J_0(x, c) = O_P(n^{-1/2}h^{-1}), \quad (24)$$

$$\sup_{x,y,h,\tau,\theta} \left| \nabla_{\theta}^2 \hat{f}_{\theta_0}^{h,\tau}(y, x) - \nabla_{\theta}^2 f_{\theta_0}^{*h,\tau}(y, x) \right| \mathbb{1}_{y \in A_{\tau}} J_{\theta}(x, c) = O_P(n^{-1/2}h^{-2}). \quad (25)$$

Furthermore, for  $x$  such as  $J_0(x, c) \neq 0$ ,

$$\begin{aligned} \left( \nabla_{\theta} \hat{f}_{\theta_0}^{h,\tau}(y, x) - \nabla_{\theta} f_{\theta_0}^{*h,\tau}(y, x) \right) &= \int \frac{\int_{\mathcal{X}} \int_t^{\tau_0} g_{f,x,y}^{h,\tau}(x_2, y_2) d\mathbb{P}(x_2, y_2) d\bar{M}^G(t)}{1 - H(t)} \\ &\quad + R_n(\tau, h, x, y), \end{aligned} \quad (26)$$

where  $\bar{M}^G(y) = n^{-1} \sum_{i=1}^n M_i^G(y)$ ,  $M_i^G$  is defined in Theorem 4,  $\sup_{x,y,\tau,h} |R_n(\tau, h, x, y)| = O_P((\log n)^{1/2} n^{-1} h^{-3})$  and  $g_{f,x,y}^h$  is defined by

$$\begin{aligned} g_{f,x_1,y_1}^{h,\tau}(x_2, y_2) &= \frac{1}{h} \frac{(x_1 - x_2) K'_h(\theta'_0 x_1 - \theta'_0 x_2) K_h(y_1 - y_2)}{f_{\theta'_0 X}^{\tau}(\theta'_0 x_1)} \\ &\quad - \frac{K_h(\theta'_0 x_1 - \theta'_0 x_2) K_h(y_1 - y_2) f_{\theta'_0 X}^{\prime\tau}(\theta'_0 x_1)}{f_{\theta'_0 X}^{\tau}(\theta'_0 x_1)^2}, \end{aligned}$$

where  $f_{\theta'_0 X}^{\prime\tau}$  denotes the derivative of  $u \rightarrow f_{\theta'_0 X}^{\tau}(u)$ , the conditional density of  $\theta'_0 X$  given  $Y \in A_{\tau}$ .

**Lemma 8.** *Under the Assumptions of Lemma 2*

$$\sup_{h,\tau} \int \frac{[\nabla_{\theta} \hat{f}_{\theta_0}^{h,\tau}(y,x) - \nabla_{\theta} f_{\theta_0}^{*h,\tau}(y,x)] \mathbb{1}_{y \in A_{\tau}} J_0(x, c/4) d\mathbb{P}(x,y)}{f_{\theta_0}^{\tau}(y, \theta'_0 x)} = o_P(n^{-1/2}).$$

*Proof of Lemma 7.* To prove (23-25), we only prove (25) since the others are similar. To prove (25), we only consider the terms in which the second derivative is involved, the others being studied analogously. Consider

$$\begin{aligned} & \frac{1}{h} \sum_{i=1}^n \delta_i W_{in}(X_i - x) K_h''(\theta' X_i - \theta' x) K_h(Z_i - y) (X_i - x)' \mathbb{1}_{Z_i \in A_{\tau}} f_{\theta' X}^{\tau}(\theta' x)^{-1} \\ &= \frac{1}{h} \sum_{i=1}^n \delta_i W_i^* J_{\theta}(X_i, c) (X_i - x) K_h''(\theta' X_i - \theta' x) K_h(Z_i - y) (X_i - x)' \mathbb{1}_{Z_i \in A_{\tau}} f_{\theta' X}^{\tau}(\theta' x)^{-1} \\ &+ \frac{1}{h} \sum_{i=1}^n \delta_i W_i^* Z_G(Z_i -) (X_i - x) K_h''(\theta' X_i - \theta' x) K_h(Z_i - y) (X_i - x)' \mathbb{1}_{Z_i \in A_{\tau}} f_{\theta' X}^{\tau}(\theta' x)^{-1}, \end{aligned}$$

where the first term is contained in  $\nabla_{\theta}^2 f_{\theta}^{*h,\tau}$ , while the second can be bounded by

$$O_P(n^{-1/2} h^{-2}) \left[ \frac{1}{nh^2} \sum_{i=1}^n \delta_i \mathbb{1}_{Z_i \leq \tau_0} |K''| \left( \frac{\theta' X_i - \theta' x}{h} \right) |K| \left( \frac{Z_i - y}{h} \right) \right].$$

Using the results of Sherman (1994), the term inside the brackets is  $O_P(1)$  uniformly in  $x, y, \theta$  and  $h$ .

Now, for the representation (26), observe that

$$\begin{aligned} & \nabla_{\theta} [\hat{f}_{\theta_0}^{h,\tau} - f_{\theta_0}^{*h,\tau}](y, x) \\ &= h^{-1} \sum_{i=1}^n \delta_i W_i^* Z_G(Z_i -) (x - X_i) K_h'(\theta'_0 x - \theta'_0 X_i) K_h(y - Z_i) f_{\theta'_0 X}^{\tau}(\theta'_0 x)^{-1} \mathbb{1}_{Z_i \in A_{\tau}} \\ &- \sum_{i=1}^n \delta_i W_i^* Z_G(Z_i -) J_0(x, c) K(\theta'_0 X_i - \theta'_0 x) K_h(Z_i - y) f_{\theta'_0 X}^{\tau}(\theta'_0 x) f_{\theta'_0 X}^{\tau}(\theta'_0 x)^{-2} \mathbb{1}_{Z_i \in A_{\tau}} \\ &+ R'_n(\tau, h, x, y), \end{aligned} \tag{27}$$

with  $\sup_{x,y,h,\tau} |R'_n(\tau, h, x, y)| = O_P(n^{-1} h^{-3/2} [\log n]^{1/2})$ , from the convergence rate of  $Z_G$  (see (21) and (22)) and the convergence rate of the denominator in (4) and its derivative, say  $(\hat{f}_{\theta'_0 X}^{\tau} - f_{\theta'_0 X}^{\tau})$  and  $(\hat{f}_{\theta'_0 X}^{\tau} - f_{\theta'_0 X}^{\tau})$  (which are of uniform rate  $O_P(n^{-1/2} h^{-1/2} [\log n]^{1/2})$  and  $O_P(n^{-1/2} h^{-3/2} [\log n]^{1/2})$  from arguments similar as for the proofs of (17)-(19) and (23)-(25)).

An i.i.d. representation of the main term in (27) can be deduced from Theorem 4 since the class  $\{h^3 g_{f,x,y}^{h,\tau}, x, y, h\}$  is a VC-class from Nolan and Pollard (1987).  $\square$

*Proof of Lemma 8.* Observe that, from classical kernel arguments

$$\sup_t \left| \iint_{x_2, t \leq y_2 \leq \tau_0} g_{f,x,y}^{h,\tau}(x_2, y_2) J_0(x, c/4) d\mathbb{P}(x_2, y_2) d\mathbb{P}(x, y) - E[\nabla_{\theta} f_{\theta_0}^{\tau}(Y, X) J_0(X, c/4)] \right| = O(h^4),$$

since  $K$  is of order 4. From the representation (26) in Lemma 7,

$$\begin{aligned}
& \int [\nabla_{\theta} \hat{f}_{\theta_0}^{h,\tau}(y, x) - \nabla_{\theta} f_{\theta_0}^{*,h,\tau}(y, x)] J_0(x, c/4) d\mathbb{P}(x, y) \\
&= \int [1 - H(t)]^{-1} E[\nabla_{\theta} f_{\theta_0}^{\tau}(Y, X) J_0(X, c/4)] d\bar{M}^G(t) \\
&\quad + \int [1 - H(t)]^{-1} \left[ \iint_{x_2, t \leq y_2 \leq \tau_0} g_{f,x,y}^{h,\tau}(x_2, y_2) J_0(x, c/4) d\mathbb{P}(x_2, y_2) d\mathbb{P}(x, y) \right. \\
&\quad \left. - E[\nabla_{\theta} f_{\theta_0}^{\tau}(Y, X) J_0(X, c/4)] \right] d\bar{M}^G(t) \\
&\quad + \int R_n(\tau, h, x, y) d\mathbb{P}(x, y), \tag{28}
\end{aligned}$$

where the last term is  $o_P(n^{-1/2})$  uniformly in  $\tau$  and  $h$ . The first one is zero from Proposition 5 and since  $J_0$  only depends on  $\theta'_0 X$ .

For the second, let  $\phi_n(t, h, \tau) =$

$$[1 - H(t)]^{-1} \left\{ \iint_{x_2, t \leq y_2 \leq \tau_0} g_{f,x,y}^{h,\tau}(x_2, y_2) J_0(x, c/4) d\mathbb{P}(x_2, y_2) d\mathbb{P}(x, y) - E[\nabla_{\theta} f_{\theta_0}^{\tau}(Y, X) J_0(X, c/4)] \right\}.$$

Using the fact that  $\mathcal{H}_n$  is of cardinality  $k_n$ , we have, for the second term in (28),

$$\mathbb{P} \left( \sup_{h \in \mathcal{H}_n} \left| \int \phi_n(t, h, \tau) d\bar{M}^G(t) \right| \geq \varepsilon \right) \leq k_n \sup_{h \in \mathcal{H}_n} \mathbb{P} \left( \left| \int \phi_n(t, h, \tau) d\bar{M}^G(t) \right| \geq \varepsilon \right).$$

Now apply Lenglart's inequality (see Lenglart (1977) or Theorem 3.4.1 in Fleming and Harrington, (1991)). This shows that, for all  $\varepsilon > 0$  and all  $\eta > 0$ ,

$$\begin{aligned}
& \mathbb{P} \left( \sup_{\tau \leq s \leq \tau_0} \left\{ \int_0^s \phi_n(t, h, \tau) d\bar{M}^G(t) \right\}^2 \geq \varepsilon^2 \right) \\
& \leq \frac{\eta}{\varepsilon^2} + \mathbb{P} \left( n^{-1} \int_0^{\tau_0} \phi_n^2(t, h, \tau) \frac{[1 - \hat{H}(t-)] dG(t)}{1 - G(t-)} \geq \eta \right). \tag{29}
\end{aligned}$$

As mentioned before,  $\sup_t |\phi_n(t, h, \tau)| = O(h^4)$ . From (29) and condition on  $k_n$  in Assumption 1, the Lemma follows.  $\square$

#### 7.4.2 Donsker classes

As stated in Assumption 7, to obtain a  $n^{-1/2}$ -convergence of  $\hat{\theta}$ , we need the regression function (and its gradient) to be sufficiently regular. In the Lemma below, we first show that the classes of functions defined in Assumption 7 are Donsker, and that  $\hat{f}_{\theta_0}^{h,\tau}$  also belongs to the same regular class as  $f_{\theta_0}^{\tau}$  with probability tending to one.

**Proposition 9.** *Consider the classes  $\mathcal{H}_1$  and  $\mathcal{H}_2$  defined in Assumption 7.  $\mathcal{H}_1$  and  $\mathcal{H}_2$  are Donsker classes. Furthermore,  $\hat{f}_{\theta_0}^{h,\tau}$  and  $\nabla_{\theta} \hat{f}_{\theta_0}^{h,\tau}$  belong respectively to  $\mathcal{H}_1$  and  $\mathcal{H}_2$  with probability tending to one for some constant  $M$  sufficiently large.*

*Proof.* The class  $\mathcal{H}_1$  is Donsker from Corollary 2.7.4 in Van der Vaart and Wellner (1996). The class  $\mathcal{H}_2$  is Donsker from a permanence property of Donsker classes, see Examples 2.10.10 and 2.10.7 in Van der Vaart and Wellner (1996). We only show the proof for  $\nabla_{\theta} \hat{f}_{\theta_0}^{h,\tau}$ , since the one for  $\hat{f}_{\theta_0}^{h,\tau}$  is similar. Write

$$\begin{aligned}
& \nabla_{\theta} \hat{f}_{\theta_0}^{h,\tau}(z, x) \\
&= \frac{1}{nh} \sum_{i=1}^n \frac{\delta_i \mathbb{1}_{Z_i \in A_{\tau}}(X_i - x) K'_h(\theta'_0 X_i - \theta'_0 x) K_h(Z_i - z) J_0(X_i, c/2)}{[1 - \hat{G}(Z_i -)] f_{\theta'_0 X}^{\tau}(\theta'_0 x)} \\
&+ \frac{1}{nh} \sum_{i=1}^n \frac{\delta_i \mathbb{1}_{Z_i \in A_{\tau}}(X_i - x) K'_h(\theta'_0 X_i - \theta'_0 x) K_h(Z_i - z) [\hat{f}_{\theta'_0 X}^{\tau}(\theta'_0 x) - f_{\theta'_0 X}^{\tau}(\theta'_0 x)]}{[1 - \hat{G}(Z_i -)] \hat{f}_{\theta'_0 X}^{\tau}(\theta'_0 x) f_{\theta'_0 X}^{\tau}(\theta'_0 x)} J_0(X_i, c/2) \\
&- \left[ \frac{1}{nh} \sum_{i=1}^n \frac{(X_i - x) K'_h(\theta'_0 X_i - \theta'_0 x) J_0(X_i, c/2)}{(f_{\theta'_0 X}^{\tau}(\theta'_0 x))^2} \right] \\
&\times \left[ \frac{1}{n} \sum_{i=1}^n \frac{\delta_i K_h(\theta'_0 X_i - \theta'_0 x) K_h(Z_i - z) \mathbb{1}_{Z_i \in A_{\tau}}}{[1 - \hat{G}(Z_i -)]} \right] \\
&+ \left[ \frac{1}{nh} \sum_{i=1}^n \frac{(X_i - x) K'_h(\theta'_0 X_i - \theta'_0 x) \left[ (\hat{f}_{\theta'_0 X}^{\tau}(\theta'_0 x))^2 - (f_{\theta'_0 X}^{\tau}(\theta'_0 x))^2 \right] J_0(X_i, c/2)}{(\hat{f}_{\theta'_0 X}^{\tau}(\theta'_0 x) f_{\theta'_0 X}^{\tau}(\theta'_0 x))^2} \right] \\
&\times \left[ \frac{1}{n} \sum_{i=1}^n \frac{\delta_i \mathbb{1}_{Z_i \in A_{\tau}} K_h(\theta'_0 X_i - \theta'_0 x) K_h(Z_i - z)}{[1 - G(Z_i -)]} \right].
\end{aligned}$$

From this expression, we clearly see that  $\nabla_{\theta} \hat{f}_{\theta_0}^{h,\tau}(y, x) = x\phi_1(x'\theta_0, y) + \phi_2(x'\theta_0, y)$ . Now we must check that  $\phi_1$  and  $\phi_2$  are in  $\mathcal{H}_1$  with probability tending to one. Since the functions are twice continuously differentiable (from the assumptions on  $K$ ), we only have to check their boundedness. From Lemma 7, this can be done at first by replacing  $\hat{f}^{h,\tau}$  by  $f^{*h,\tau}$  (i.e.  $\hat{G}$  by the true function  $G$ ). Among the several terms in the decomposition of  $\nabla_{\theta} f^{*h,\tau}$ , we will only study

$$\phi(u, y) = \frac{1}{nh} \sum_{i=1}^n \frac{\delta_i \mathbb{1}_{Z_i \in A_{\tau}} X_i K'_h(\theta'_0 X_i - u) K_h(Z_i - z) J_0(X_i, c/2)}{[1 - G(Z_i -)] f_{\theta'_0 X}^{\tau}(u)},$$

since the others are similar. We will show that the derivatives of order 0, 1 and  $1 + \delta$  of this function are uniformly bounded by some constant  $M$  with probability tending to one.

Now a centered version of  $\phi$  converges to zero at rate  $O_P([\log n]^{1/2} n^{-1/2} h^{-1})$  (see Einmahl and Mason (2005)), which tends to zero as long as  $nh^2 \rightarrow \infty$ . Furthermore,  $E[\phi]$  is uniformly bounded from our Assumption 7 on the regression function. For the derivative,

$$\begin{aligned}
\partial_u \phi(u, y) &= -\frac{1}{nh} \sum_{i=1}^n \frac{\delta_i \mathbb{1}_{Z_i \in A_{\tau}} X_i K''_h(\theta'_0 X_i - u) K_h(Z_i - z) J_0(X_i, c/4)}{[1 - G(Z_i -)] f_{\theta'_0 X}^{\tau}(u)} \\
&- \frac{1}{nh} \sum_{i=1}^n \frac{\delta_i \mathbb{1}_{Z_i \in A_{\tau}} X_i K'_h(\theta'_0 X_i - u) K_h(Z_i - z) J_0(X_i, c/4) f_{\theta'_0 X}^{\tau}(u)}{[1 - G(Z_i -)] (f_{\theta'_0 X}^{\tau}(u))^2}.
\end{aligned}$$

Again,  $E[\partial_u \phi]$  is uniformly bounded from our Assumption 7. Now using the results of Einmahl and Mason (2005), the centered version of  $\partial_u \phi$  tends to zero provided that  $nh^6 \rightarrow \infty$ . The same arguments apply for  $\partial_y \phi$ . Hence, with  $f_i(u, y) = E(\phi_i(u, y))$  we proved that  $\sup_{u, y} |\partial_u^j \partial_y^k \phi_i(u, y) - \partial_u^j \partial_y^k f_i(u, y)|$  tends to zero in probability for  $i = 1, 2$ ,  $k + j \leq 1$ . Now we have to show that  $\partial_u \phi_j$  and  $\partial_y \phi_j$  are  $\delta$ -Hölder for  $j = 1, 2$  with an Hölderian constant bounded by some  $M$  with probability tending to one. We only prove the result for  $\partial_u \phi_1$ . We have

$$\begin{aligned} \sup_{u', y', x, y} \frac{|\partial_u \phi_1(u, y) - \partial_u \phi_1(u', y')|}{\|(u, y) - (u', y')\|^\delta} &= \max \left( \sup_{|u-u'| \geq n^{-1}, y, y'} \frac{|\partial_u \phi_1(u, y) - \partial_u \phi_2(u', y')|}{\|(u, y) - (u', y')\|^\delta}, \right. \\ &\quad \left. \sup_{|u-u'| \leq n^{-1}, y, y'} \frac{|\partial_u \phi_1(u, y) - \partial_u \phi_1(u', y')|}{\|(u, y) - (u', y')\|^\delta} \right) \\ &= \max(S_1, S_2). \end{aligned}$$

We have

$$\begin{aligned} S_1 &\leq \sup_{u, y, u', y'} \frac{|\partial_u f_1(u', y') - \partial_u f_1(u, y)|}{\|(u', y') - (u, y)\|^\delta} \\ &\quad + 2n^\delta \sup_{u, y, u', y'} |\partial_u \phi_1(u, y) - \partial_u f_1(u, y)|. \end{aligned}$$

From our Assumptions, the first supremum is bounded, while the last is

$O_P(n^{-1/2+\delta}[\log n]^{1/2}h^{-3})$  from the convergence rate of  $\partial_u \phi_2$ . It tends to zero provided that  $nh^{6+\delta} \rightarrow \infty$ . For  $S_2$ , since  $K$  is  $\mathcal{C}^3$  with bounded derivatives, for some positive constant  $M$ ,

$$\begin{aligned} \sup_{\|(u, y) - (u', y')\| \leq n^{-1}, y, y'} \frac{|\partial_u \phi_1(u, y) - \partial_u \phi_1(u', y')|}{\|(u, y) - (u', y')\|^\delta} &\leq M \times O_P(1) \left\| \sum_{i=1}^3 K^{(i)} \right\|_\infty \\ &\quad \times \sup_{\|(u, y) - (u', y')\| \leq n^{-1}} \|(u, y) - (u', y')\|^{1-\delta} h^{-1} \frac{1}{nh^4} \sum_{i=1}^n \frac{\delta_i}{1 - G(Z_i -)}. \end{aligned}$$

The last supremum is bounded by  $O_P(1) \times n^{-1+\delta}h^{-5}$ , and it tends to zero when  $nh^6 \rightarrow \infty$  (and the  $O_P(1)$  term does not depend on  $u, y$ ).  $\square$

## References

- Bashtannyk, D. M. and Hyndman, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Comput. Statist. Data Anal.* **36**(3), 279–298.
- Bosq, D. and Lecoutre, J.P. (1997). *Théorie de l'estimation fonctionnelle*. Vol. 3 of *Economie et statistiques avancées*. Economica, Paris.

- Brunel, E. and Comte, F. (2006). Adaptive nonparametric regression estimation in presence of right censoring. *Math. Methods Statist.* **15**(3), 233–255.
- Cox, D. R. (1972). Regression models and life-tables, *J. Roy. Statist. Soc. Ser. B* **34**, 187–220.
- Delecroix, M., Hristache, M. and Patilea V. (2006). On semiparametric  $M$ -estimation in single-index regression. *J. Statist. Plann. Inference* **136**(3), 730–769.
- Delecroix, M., Lopez, O. and Patilea V. (2008). Nonlinear censored regression using synthetic data. *Scand. J. Statist.* **35**(2), 248–265.
- Delecroix, M., Härdle W. and Hristache M. (2003). Efficient estimation in conditional single-index regression. *J. Multivariate Anal.* **86**(2), 213–226.
- Dominitz, J. and Sherman R. P. (2005). Some convergence theory for iterative estimation procedures with an application to semiparametric estimation. *Econometric Theory* **21**(4), 838–863.
- Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.* **33**(3), 1380–1403.
- Fan, J. and Yim, T. H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika* **91**(4), 819–834.
- Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York.
- Gannoun, A., Saracco, J., Yuan, A. and Bonney, G. E. (2005). Non-parametric quantile regression with censored data. *Scand. J. Statist.* **32**(4), 527–550.
- Gill, R. (1983). Large sample behaviour of the product-limit estimator on the whole line. *Ann. Statist.* **11**(1), 49–58.
- Heuchenne, C. and van Keilegom I. (2007). Nonlinear regression with censored data. *Technometrics* **49**(1), 34–44.
- Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58**(1-2), 71–120.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457–481.

- Koul, H., Susarla V. and van Ryzin J. (1981). Regression analysis with randomly right-censored data. *Ann. Statist.* **9**(6), 1276–1288.
- Lopez, O. (2008). Single-index regression models with right-censored responses. *To appear in J. Statist. Plann. Inference* .
- Lopez, O. and Valentin P. (2008). Nonparametric lack-of-fit tests for parametric mean-regression models with censored data. *To appear in J. Multivariate Anal.* .
- Lu, X. and Burke M. D. (2005). Censored multiple regression by the method of average derivatives. *J. Multivariate Anal.* **95**(1), 182–205.
- Miller, R. and Halpern J. (1982). Regression with censored data. *Biometrika* **69**(3), 521–531.
- Nolan, D. and Pollard D. (1987).  $U$ -processes: rates of convergence. *Ann. Statist.* **15**(2), 780–799.
- Sánchez Sellero, C., González Manteiga, W. and Van Keilegom, I. (2005). Uniform representation of product-limit integrals with applications. *Scand. J. Statist.* **32**(4), 563–581.
- Satten, G. A. and Datta S. (2001). The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. *Amer. Statist.* **55**(3), 207–210.
- Sherman, R. P. (1994). Maximal inequalities for degenerate  $U$ -processes with applications to optimization estimators. *Ann. Statist.* **22**(1), 439–459.
- Stute, W., González Manteiga, W. and Sánchez Sellero, C. (2000). Nonparametric model checks in censored regression. *Comm. Statist. Theory Methods* **29**(7), 1611–1629.
- Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *J. Multivariate Anal.* **45**(1), 89–103.
- Stute, W. (1995). The central limit theorem under random censorship. *Ann. Statist.* **23**(2), 422–439.
- Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scand. J. Statist.* **23**(4), 461–471.
- Stute, W. (1999). Nonlinear censored regression. *Statist. Sinica* **9**(4), 1089–1102.
- Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probab.* **22**(1), 28–76.

- Van der Vaart, A. W. (1998). *Asymptotic statistics*, Vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.
- Van der Vaart, A. W. and Wellner J. A. (1996). *Weak convergence and empirical processes*, Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.
- Wei, L. J., Ying Z. and Lin D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77**(4), 845–851.
- Zhou, M. (1992). Asymptotic normality of the “synthetic data” regression estimator for censored survival data. *Ann. Statist.* **20**(2), 1002–1021.